



The 7th International Conference on Information Technology

Big Data

ISSN 2306-6105



ISBN 978-9957-8583-3-9

Conference Proceeding

Full

Prepared and Edited by:

- ICIT15 General Chair
 - Ali Al-Dahoud, Dean of Science and IT Faculty
- Editorial Board
 - Hani Mimi, ICIT15 Co-chair
 - Khalid Jaber, ICIT5 Co-chair
 - Israa Sabatin, Designer
 - Hanade Al-Shawabkeh, Editor
 - Ayman Al-Qafa'an, Editor

The Hashemite Kingdom of Jordan

The Deposit Number at the National Library

(2015/4/1450)

009

نسخة / مركز
الإيداع

Information Technology (7:Amman:2015)

The 7th International Conference on Information Technology /
Ali As'ad Al-Dahoud. – Amman: Al-Zaytoonah University of
Jordan, 2015

(163) p.

Deposit No. : 2015/4/1450

Descriptions: /Computers //Conferences//Information
Technology/

يتحمل المؤلف كامل المسؤولية القانونية عن محتوى مصنفه ولا يعبر هذا المصنف
عن رأي دائرة المكتبة الوطنية أو أي جهة حكومية أخرى.

Under the patronage of her Excellency
The Minister of Information and Communications Technology

Al-Zaytoonah University of Jordan was established in 1993. Since then, Al-Zaytoonah has witnessed considerable progress, at both the infrastructure and academic levels. It now includes seven faculties, encompassing 28 undergraduate programs and 5 graduate programs.

Al-Zaytoonah is now a member of the following associations:

- Federation of Arab Universities
- Union of Arab & European Universities
- International Association of University (IAU)
- Federation of the Universities of the Islamic World (FUIW)
- Association of Private Institutions for Higher Education

The 7th International Conference on Information Technology

The **ICIT 2015** is indexed by EBSCO, Google scholar, ULRICHSWEB, and IET Inspec

The International Conference on Information Technology, **ICIT 2015**, is held every 2 years since 2003. This year it is the 7th conference. **ICIT 2015** is a forum for scientists, engineers, and practitioners to present their latest research results, ideas, developments, and applications in all areas of Information Technology. **ICIT 2015** will include presentations of contributed papers and state-of-the-art lectures by invited keynote speakers. Moreover, the program will include tutorials on hot areas of Information Technology. All submitted papers will go through double-blind reviewing processes by at least three reviewers. Extended versions of the conference best selected papers will be evaluated to be published in international journals that are announced each time the conference is held. Moreover, Accepted papers will be assigned a unique URL and a DOI.

ICIT 2015 has many features that make it distinguished conference among other international conferences. One of the most important features is that it is a non-profit conference, which means that the registration is free. In addition, one author of each accepted paper usually enjoys a full board hosting in Jordan during the conference for 3 nights /4 days.

Contents

Contents	I
Preface : General Chair	X
Conference Committees	XII
Co-Program Chairs	XII
Steering Committee	XII
International and Local Program Committee	XIII
International Advisory Committee	xv
Editing Committee	xvi
Technical Program Committee	xvii
Countries	xviii
Keynote Speakers	xix
Exploiting Artificial Intelligence Technology in e-Health	xix
Security in Cloud Computing	xx
Applied Visual Computing: Challenges and Opportunities	xxi
What is after image processing!	xxii
Workshops	xxiii
Workshop on Artificial Intelligence in e-Learning and Education	xxiii
Consumer Electronics Information Technologies in 21st Century Course or Blessing!	xxiv

Artificial Intelligence.....	
Using temporal formalisms to support organizational creativity.....	1
Multi Objects Tracking in Nighttime Traffic Scenes	8
New Selection Schemes for Particle Swarm Optimization	17
Knowledge Acquisition for Developing Knowledge-Base of Diabetic Expert System	26
Solving Nurse Rostering Problem Using Artificial Bee Colony Algorithm.....	32
Towards Developing an Intelligent HAJJ Guide system Pilgrim Tracking and Identification Using Mobile Phones	39
Online Recognition System for Handwritten Arabic Digits.....	45
Hierarchical Singular Value Decomposition for Halftone Images	50
The Rise of the Robotic Judge in Modern Court Proceedings	59
Design of rectangular microstrip antenna with rectangular aperture in the ground plane using artificial neural networks	68
Design and Implementation of Two Degree of freedom Proportional Integral Derivative Controller	75
Experiments with Simulated Humanoid Robots.....	81
Dynamic Frames-Based Generation of Web 2.0 Applications.....	90
Unsupervised Classification of Mobile Device Images	96
Distributed 3D Object Recognition System Using Smartphones.....	102
A Survey on Digital Image Steganography.....	109
Computer Vision Applied to Road Lines Recognition Using Machine Learning	116

Combining ICA Representations for Recognizing Faces	122
Modeling and Design of Anisotropic Circular Microstrip Patch Antenna Using Neurospectral Computation Approach	127
Bridging the Gap between Modeling of Mobile Agent-based Systems and Semantic Web using Meta-Modeling and Graph Grammars	134
Artificial Bee Colony Based Focus Fusion	142
Ontology-based knowledge recognition in service-oriented virtual research environments Case: application in e-learning	148
Particle Swarm Optimization Based Discrete Cosine Transform for Person Identification by Gait Recognition	156
Supporting Arabic Sign Language Recognition with Facial Expressions	164
Bioinformatics and Computational Biology	
Comparison between X-Ray Radiography Image Fusion Algorithms Used in Medical Applications	171
Detecting patients with Parkinson’s disease using PLP and VQ	177
Software for Reaction-Time Measurement and its Application for the Evaluation of Patient's Recovery after the Stroke	182
Cloud Computing	
SwiftEnc: Hybrid Cryptosystem with Hash-Based Dynamic Key Encryption	186
Proposed Evaluation Criteria for Selecting Appropriate Cloud Based On-Demand CRM for SMEs.	193
Ensuring Smart Grid Data Security at Cloud Data Centres	201
Consistency tradeoffs on distributed multi-datacentric systems SCOLCH theorems	208

Computer Networks and Communications	
Analyzing Optimal Setting Of Reference Point Group Mobility Model Using DSR Protocol In MANETS	213
Quantum information technology: novel way for increase of sensory systems capability.....	223
The effect of using channel equalizer in the SDR Modem	232
Cost Efficient Fast Autonomous Reconfiguration System in Wireless Mesh Networks	240
Parameter Analysis for Clustering in Manet in Disaster Scenarios	247
Extending OpenFlow in Virtual Networks	252
An Overview of Integration of Mobile Infrastructure with SDN/NFV Networks...	259
Impact of Node Clustering on Power Consumption in WSN A Comparative Study	266
Application Layer Protocols to Protect Electronic Mail from Security Threads	270
Frequency Assignment for Cellular Mobile Systems Using a Memetic Algorithm	275
WSN for AIR quality Monitoring in Annaba City.....	283
Complex Adaptive WSNs for Polluted Environment Monitoring	289
Computers and Networks Security.....	
Enhancing Intrusion Detection System (IDS) by Using Honeybee Concepts and Framework.....	297
On The Improvement of the Tri-Way Pixel Value Differencing Steganography Algorithm	303
Secure Data Sharing Polices and Architecture Preserving Privacy	308

Groebner Bases and Coding	315
Hide image in image based on LSB Replacement and Arnold Transform	319
Finger-Knuckle-Print identification System Using Hidden Markov Model and Discret Cosine Transform.....	324
A Suggested Algorithm of Recommender System to Recommend crawled-Web Open Educational Resources to Course Management System	330
Temperature Aware Design for High Performance Processors.....	338
Image Watermarking using DC Component of DCT	345
A New Model for Pre-analysis of Network Traffic Using Similarity Measurement	349
A New Authenticated Key Agreement Protocol	354
The Generalised Secured Mobile Payment System Based on ECIES and ECDSA...	359
Database and Data Mining	
A Formal Mathematical Semantics of Advanced Operations of Multiset Table Algebra.....	369
Temporal Data in Enterprise Database Systems	376
Analysis of Oral Cancer Prediction using Features Selection with Machine Learning	383
Decision Support System, Utilizing Data Warehouse Technique For the Tourism Sector in Egypt.....	389
Web Crawler System for Distinct Author Identification in Bibliographic Databases	295
Technical object as a system of “growing” structure	303
E-Technology	

Novel Review Of Electronic Government Stages Among Different Continents	309
Predictors of Mobile Learning Adoption The Case of Universiti Teknologi MARA	317
Predictors of E-Participation Levels: The Case of Jordan	325
The Development of Software Agents in e-Learning 3.0.....	332
Privacy Policy of E-Government Websites and the Effect on Users' Privacy	338
E-Government Adoption in Jordan: The Influence of Age.....	345
Big Issues for A Small Piece: RFID Ethical Issues.....	351
The Effect of Using Social Media in Governments: Framework of Communication Success.....	357
A Review on Internet Banking Security and Privacy Issues in Oman.....	365
Information and Knowledge Engineering	
A survey on Applying Ontological Engineering Approach for Hepatobiliary System Diseases	370
Hierarchical Sparsity-Regularized Framework Based Frequency Hopping Spectrum Estimation With Antenna Array System	376
Unsupervised Single Channel Source Separation with Nonnegative Matrix Factorization	382
The Trends and Directions of Wisdom and Semantic-based Search System	386
Evaluating the Success of Information Strategic System Planning (Two Cases from Jordan)	390
A Novel Web Application for Image Fusion.....	397
Ontology-based Facilitation Support Tool for Group Decision Making.....	402
Quality Driven Approach for Data Integration Systems	409

Internet and Web Services	
Privacy and Protection in Electronic Transaction: A Review of The E-Commerce Security Protocols.....	421
Towards Building Novel Educational System for School Students Using Smart Phones and QR Codes.....	430
Cross-Language Name Matching for Data Fusion in Linked Open Data.....	436
Exploring Domain Interrelations in Freebase Schema Using Modularity-Based Community Detection	443
The Arabic Language Status in the Jordanian Social Networking and Mobile Phone Communications.....	449
Similar Searching, Unsimilar Clicking.....	457
A Query Log-Based Study of Cross-Nation Perception.....	463
Internet-based Troubleshooting and Monitoring System of Industrial Robots	470
Modeling and Simulations	
Multicore RISC Processor Implementation by VHDL for Educational Purposes....	476
Numerical Simulation of Axial Coolant Flow in Rod Bundles of a Nuclear Reactor	484
VeSimulator A Location-Based Vehicle Simulator Model for IoT Applications	490
Numerical Simulation for Fuzzy Fredholm Integral Equations Using Reproducing Kernel Algorithm.....	497
Modeling and Simulation of Electroactive Polymer Robotic Actuator.....	502
Introduction of Modeling Complex Management Systems using Fuzzy Cognitive Map.....	508

UML Activity Diagrams and Maude Integrated Modeling and Analysis Approach Using Graph Transformation	515
Particle swarm optimization and method of moments for modeling and optimization of microstrip antennas	522
Simulation of Class D resonance inverter for Acoustics Energy Transfer applications	527
Symbolic Modeling Approach in Verification and Testing.....	533
Design, Implementation and Comparison of Low-Cost Laser Scanning Systems for 3D Modeling	540
Plane Segmentation of Kinect Point Clouds using RANSAC.....	545
Instruments of Operator’s Active State Identification	552
Multimedia and Its Applications.....	
How Infographic should be evaluated?.....	558
Noise Reduction Algorithm Based on Complex Wavelet Transform of Digital Gamma Ray Spectroscopy	565
Designing Children’s Encyclopedia (3D Dinosaur) Via Augmented Reality Marker-Based Interaction.....	571
Human Activity Recognition for Surveillance Applications	577
A Comparative Study on The Existing Graphical User Interfaces for Occupational Therapy.....	587
New Technique of Forensic Analysis for Digital Cameras in Mobile Devices	597
Virtual Tourism Application through 3D Walkthrough: Flor De La Mar	603
Enhanced Watermarking Scheme for 3D Mesh Models	612

User interfaces applied to teleoperate mobile robots with keyboard command, PS3 controller and mobile phone 620

Vowels as HCI for Controlling Mouse Cursor 625

Parallel and Distributed Systems 625

 An Effective Parallel FDTD Algorithm For Modeling 3D Frequency-Dependent Electromagnetic Applications 632

 The Dualism of Context in Ubiquitous Computing 637

 Anatomy of the Parallel Tree Based Strategy for High Strength Interaction Testing 643

Software Engineering 643

 Using MADA+TOKAN to Generate Use Case Models from Arabic User Requirements in a Semi-Automated Approach..... 652

 Smart OptiSelect Preference Based Innovative Framework for User-in-the-Loop Feature Selection in Software Product Lines..... 657

 Decision Support System for Learning Disabilities Children in Detecting Visual-Auditory-Kinesthetic Learning Style 667

 Identification of Potential Crime Tactical Path-Finding Using Analytical Hierarchy Process (AHP) in Situational Crime Prevention : Crime Intelligence in New Era... 672

 Extended Cavity Model to Analysis Tunable Circular Disk Microstrip Antenna Using Genetic Algorithm..... 679

 Improving the Reuse of Services in Geospatial Applications with XMDD Technology 685

 A Domain-Specific Language for Service Level Agreement Specification 693

 Anatomy of the Tree Based Strategy for High Strength Interaction Testing..... 698

Preface : General Chair

The 7th International Conference on Information Technology

ICIT 2015
Big Data
ISSN 2306-6105
ISBN 978-9957-8583-3-9



Welcome to ICIT'15; the 7th International Conference on Information Technology. ICIT'15 has attracted high quality papers in different fields of IT. It offers a unique opportunity for Arab scientists and practitioners to meet and get in touch with outstanding international scientists to share their expertise, research results, and achievements. During the last decade, Jordan has made IT fundamental to its development. This conference is aimed to promote this technology nationally and internationally. ICIT'15 get indexed and cosponsored by IEEE, EBSCO, IET Inspec, ULRICHS, and Google scholar. Each paper has been blind reviewed by three reviewers, two reviewers from our international committee (from 45 countries) and one reviewer from the conference local committee. This year, we were delighted to have 302 papers submitted from 53 different countries, of these, 120 were accepted for presentation at the conference with an acceptance rate of 39%.

Four distinguished keynote speakers will feature lectures during the conference's three-day alongside with two workshops that will be presented this year. Prof. Anu Gokhale; Illinois State University, USA; will share his experience in the field of "Cloud Computing Security". Prof. Hassan Ugail; Director of Centre for Visual Computing, University of Bradford, UK; in his speech will present the recent developments on "Applied Visual Computing: Challenges and

Opportunities". Furthermore, Prof. Abdel-Badeeh Salem, Head of Artificial Intelligence and Knowledge Engineering Department, Ain Shams University, Egypt; will introduce the latest trends in "Exploiting Artificial Intelligence Technology in e-Health". Finally; in Dr. Mohamed Kayyali (Ongoing for Harvard University) talk, he will provide an answer for the question "What is after Image Processing!".

Many individuals have contributed to the success of this event. My sincere gratitude is addressed to the authors who chose to submit their work to the ICIT'15, as well as to the reviewers and all conference committees. Your valuable time and effort are highly appreciated. I also express my deep appreciation to our workshops organizers; Sheik Ali Alao (ePromaG Consultancy London, UK) and Prof Abdel-Badeeh Salem, who elicited great workshop proposals on important and timely topics.

Moreover, I would like to thank the editors of all journals that agreed to publish the extended versions of our conference best accepted papers in special or regular issues of their journals. Besides that, I wish to acknowledge the outstanding work put in over many months by the staff of Science & IT faculty at Al-Zaytoonah University of Jordan, who contributed their expertise and time generously to making the conference a success, specially the vice Dean and the Department's Chairs. And I would like to single out, Dr. Hani Mimi, Dr Khalid Jaber, and Miss Israa Al-Sabatin for their remarkable assistant.

The ICIT'15 would not have seen the light without the encouragement and full support from our university president Prof. Rushdi A. Hasan and the board of directors.

In closing, on behalf of all committees of the ICIT'15, I wish you a fruitful conference and pleasant stay in Jordan.

Al-Dahoud Ali; SMIEEE, ACM
ICIT'15; General Chair,
Dean of Science and IT faculty,
Al-Zaytoonah University of Jordan
Aldahoud@zuj.edu.jo

Conference Committees

Co-Program Chairs

- Hani Mimi, Al-Zaytoonah University of Jordan
- Khalid Jaber, Al-Zaytoonah University of Jordan
- Babatunde Ali Alao, ePromaG Consultancy London, UK

Steering Committee

- Abdel-Badeeh Salem , Ain Shams University- Egypt
- Ahmad Al-Khasawneh, Hashemite University- Jordan
- Ahmad Dalal'ah, , University of Ha'il- Saudi Arabia
- Amjad RATTROUT , University Lyone1- France
- Anne-Marie Di Sciallo , University of Quebec- Canada
- Babatunde Ali Alao , ePromaG Consultancy London- United Kingdom
- Fredrick Japhet Mtenzi , Dublin Institute of technology- Ireland
- George S. Oreku, TIRDO/ North West University- South Africa
- Hans-Dieter Burkhard , Humboldt University Berlin- Germany
- Ion Tutanescu , University of Pitești- Romania
- Ismail Ababneh , AL El Bayt University - Jordan
- Jyoti Singhai , MANIT university- India
- Marina Santini , Uppsala University- Sweden
- Mirjana Ivanović , University of Novi Sad
- Mohamad Qatawneh , Jordan University- Jordan
- Mohamed Fezari , Badji Mokhtar Annaba University- Algeria
- Mohamed I. Roushdy , Ain Shams University- Egypt
- Mohamed JEMNI , University of TUNIS
- Rehab M. Duwairi, JUST- Jordan
- Sonja Restic , University of Novi Sad

- Valérie Monfort , Univ. Paris1 Panthéon Sorbonne- France
- Walid Salameh , PSUT- Jordan
- Wojciech Zamojski , Wrocław University of Technology- Poland
- Yukako Yagi , Harvard Medical School, Boston, MA- United States

International and Local Program Committee

- Abdallah Hlayel, Jordan
- Abdallah Qusef, Jordan
- Abdelfatah Tamimi, Jordan
- Adnan Asber, Jordan
- Adnan Hnaif, Jordan
- Adriana Carniello, Brazil
- Afaf Al-Neaimi, Iraq
- Afif Almgawish, Syria
- Ahmad Mazhar , Jordan
- Ahmad Abusukhon, Jordan
- Ahmad Alkhatib, Jordan
- Ahmad Al-Qerem, Jordan
- Ahmad Althunibat, Jordan
- Akram Abdelqader, Jordan
- Alaa Al-Afeef, United Kingdom
- Aldo Pardo, Colombia
- Alejandra Guadalupe Silva Trujillo,
- Amer Abu Salem, Jordan
- Amir Shafie, Malaysia
- Amnah El-Obaid, Jordan
- Ana Lucila Sandoval Orozco,
- Andreia Carniello, Brazil
- Antoanela Naaji, Romania
- Ashim Saha, India
- Awat Al-Nakshabendi, Iraq
- Ayat Jaradat, Jordan
- Ayman Abdalla, Jordan
- Banan Maayah, Jordan
- Baraah Alsaq, Jordan
- Basem Alokush, Jordan
- Bassam Naji Al-Tamimi, Saudi Arabia
- Benmohammed Mohamed, Algeria
- Bilal Hawashin, Jordan
- Carla Salazar Serrudo, Bolivia
- Carlos Avalos,
- Chadi Zammar, Kuwait
- Cherif Foudil, Algeria
- Dara Aqel, Jordan
- Dmytro Bui, Ukraine
- Edward Jaser, Jordan
- El-Sayed M. El-Horbaty, Egypt
- Elzbieta Wyslocka, Poland
- Erdem Yazgan, Turkey
- Eyas Elqawasmeh,

- Fadel Altamimi, Jordan
- Farhan Abdulfattah, Jordan
- Fatiha Merazka, Algeria
- Fatima Riouch, Malaysia
- Feras Al-Tarawneh, Jordan
- Fuad El-Qirem, Jordan
- Ghassan Samara, Jordan
- Gordana Milosavljevic, Serbia
- Hamid Nebdi, Morocco
- Hani Al-Mimi, Jordan
- Harith A. Dawood, Iraq
- Heider Wahsheh, Saudi Arabia
- Hikmat Abdullah, Iraq
- Irma Aslanishvili, Georgia
- Issa Otoum, Jordan
- Ivan Luković, Serbia
- Jamil Itmazi, Palestine
- Jan Szybka, Poland
- Jeanne Schreurs, Belgium
- Kateryna Solovyova,
- Khalid Farhan, Jordan
- Khalid Mohammad Jaber, Jordan
- Khalil Awad, Jordan
- Khulood Abu Maria, Jordan
- Klimis Ntalianis, Greece
- M. A. H. Akhand, Bangladesh
- Mahendra Pratap Singh, India
- Maher Nabelsi, Jordan
- Majid Jawad, Iraq
- Malik Fakri, Saudi Arabia
- Mario Pepur, Croatia
- Marzooq Al-Maitah, Jordan
- Mihaela Badea, Romania
- Mirela Erić, Serbia
- Mirjana Ivanovic, Serbia
- Mohamed Roushdy, Egypt
- Mohamed-Khireddine, Algeria
- Mohammad Abdallah, Jordan
- Mohammad Al Rawajbeh, Jordan
- Mohammad Rasmi Al-Mousa, Jordan
- Mohammed Basher, Saudi Arabia
- Mohammed Alia, Jordan
- Mohammed Elbes, Jordan
- Mohammed Lafi, Jordan
- Mohd Khaled Shambour, Jordan
- Mokhtar Kerwad, Libya
- Mosa Salah, Jordan
- Mostafa Abdel Aziem, Egypt
- Muhammed Ibrahim, Iraq
- Muhannad Abu-Hahsem, Jordan
- Munadil K. Faaeq, Iraq
- Mustafa Alrifaae, Jordan
- Mustafa Hammad, Jordan
- Nabil Arman, Palestine
- Nada Al Sallami, Jordan
- Nagham Al-Madi, Jordan

- Nancy Al Ramahi, Jordan
- Nebal Aljamal, Jordan
- Nesreen Hamad, Jordan
- Nina Rizun, Ukraine
- Nirmalya Kar, India
- Okba Kazar, Algeria
- Omaima Al-Allaf, Jordan
- Osama Alia, Saudi Arabia
- Phan Cong Vinh, Viet Nam
- Qasem Obeidat, Jordan
- Rana Bader, Jordan
- Sabarina Ismail, Malaysia
- Saidatul Norlyana Azemi, Malaysia
- Seifedine Kadry, Kuwait
- Shadi Alzubi, Jordan
- Shrouq Shihadeh, Jordan
- Siham Hamadah, Jordan
- Siva Rama Krishna Sakshi, India
- Smaranda Belciug, Romania
- Sokyna Al-Qatawneh, Jordan
- Sonja Ristic, Serbia
- Srdjan Skrbic, Serbia
- Subhrajyoti Deb, India
- Subramanian Ls, India
- Taha Elarif, Egypt
- Thamer Alrawashdeh, Jordan
- Thomas Schramm, Germany
- Tomasz Petech-Pilichowski, Poland
- Tomasz Walkowiak, Poland
- Tulshi Bezboruah, India
- Udo Averweg, South Africa
- Vaidas Giedrimas, Lithuania
- Vitaliy Mezhujev, Malaysia
- Vladimir Sulov, Bulgaria
- Yousra Harb, United States
- Zeyad Mohammad, Jordan
- Zine Eddine Bouras, Algeria

International Advisory Committee

- Anna Sołtysik-Piorunkiewicz , University of Economics in Katowice- Poland
- Atis Kapenieks , Director of Riga Technical University, Distance Education Study Centre- Latvia
- Celina Olszak Katowice , University of Economics- Poland
- Eleni C. Gkika , University of Applied Sciences
- Fatiha Merazka, University of Science & Technology Houari Boumediene - Algeria
- George S. Oreku, TIRDO/ North West University- South Africa
- Halina Kwasnicka , Wroclaw University of Technology- Poland

- Hassan Ghazal , University Mohammed Premier- Morocco
- Iman Osta , Lebanese American University- Lebanon
- Ivan Luković, University of Novi Sad- Serbia
- Jeanne Schreurs , Emeritus of Hasselt University in Belgium- Belgium
- Joan lo , University of Huddersfield
- Klimis Ntalianis, Technological Educational Institute of Athens- Greece
- Krassimir Markov , president of ITHEA International Scientific Society- Bulgaria
- Liliana M. Moga , Dunarea de Jos University of Galati- Romania
- Maria Mach-Król Katowice , University of Economics- Poland
- Michael Gr. Voskoglou , Graduate Technological Educational Institute of Patras- Greece
- Mohamed Ismail Roushdy , Ain Shams University- Egypt
- Professor Dorota Jelonek , Czestochowa University of Technology- Poland
- Roumen Kountchev , Technical University of Sofia- Bulgaria
- Roumiana Kountcheva , T&K Engineering, Sofia- Bulgaria
- Sarma Cakula , Vidzeme University of Applied Sciences- Latvia
- Štefan Korečko , Technical University of Košice
- Thomas Risse , Zentrum für Informatik und Medientechnologien- Germany
- Urszula Markowska-Kaczmar , Wroclaw University of Technology- Poland
- V. Rajamani , Anna University- India
- V.Parthasarathy , Vel Tech Multi Tech Dr Rangarajan Dr Sakunthala Engineering College- India
- Vladimir Romanov , Institute of Cybernetics of National academy of sciences of Ukraine- Ukraine
- Yukako Yagi , Harvard Medical School, Boston, MA- United States

Editing Committee

- Feras Alazzeah, Quality Assurance Office - ZUJ- Jordan

- Khalil Awad, Conference Webmaster - ZUJ- Jordan
- Israa Al-Sabatin, Conference Designer- Jordan

Technical Program Committee

- Antoanela Naaji , "Vasile Goldis" Western University- Romania
- Christina Siontorou , University of Piraeus- Greece
- Ala Khalifa, IEEE Jordan Section communication chapter - German Jordan University- Jordan
- Ali Maqousi, IEEE Jordan Section Secretary-Petra University- Jordan
- Edward Jaser, IEEE Jordan Section computer chapter chair-Princess Sumaya University of Technology- Jordan
- Elena Nechita , Vasile Alecsandri University- Romania
- Elzbieta Wyslocka , Czestochowa University of Technology- Poland
- Eva Milková , Faculty of Science - University of Hradec Králové- Czech Republic
- Joan Lu , University of Huddersfield- United Kingdom
- Kostagiolas Petros , Ionian University- Greece
- Malgorzata Pankowska , University of Economics in Katowice- Poland
- Nadia Baeshen , University of Business and Technology- Saudi Arabia
- Nina Rizun , Alfred Nobel University- Ukraine
- Raja Fenniche , Université de la Manouba- Tunisia
- Rajamani Vayanaperumal , Vel Tech Multi Tech Dr Rangarajan Dr Sakunthala Engineering College- India
- Rajendran Periyasamy , Knowledge Institute of Technology- India
- Tatiana Hrivíková , University of Economics in Bratislava- Slovakia

Countries

1. Algeria
3. Argentina
5. Austria
7. Bahrain
9. Belarus
11. Bosnia
13. Bulgaria
15. Cameroon
17. China
19. Croatia
21. Czechoslovakia
23. Denmark
25. Ecuador
27. Egypt
29. Georgia
31. Germany
33. Greece
35. Hungary
37. India
39. Iran
41. Iraq
43. Ireland
45. Italy
47. Jordan
49. Latvia
51. Libya
53. Macedonia
2. Malaysia
4. Mexico
6. Montenegro
8. Morocco
10. Nepal
12. Netherlands
14. Nigeria
16. Oman
18. Pakistan
20. Palestine
22. Poland
24. Qatar
26. Russia
28. Saudi Arabia
30. Serbia
32. Slovakia
34. Spain
36. Sudan
38. Thailand
40. Tunisia
42. Turkey
44. UK
46. Ukraine
48. Uruguay
50. USA
52. Venezuela

Keynote Speakers

EXPLOITING ARTIFICIAL INTELLIGENCE TECHNOLOGY IN E-HEALTH

Prof Dr. Abdel-Badeeh M. Salem

Head of Artificial Intelligence and Knowledge Engineering Research Labs, AinShams University,
 Ain Shams University, Cairo-Egypt

Abstract



Artificial intelligence (AI) techniques have been proved to be effective and efficient in developing intelligent systems for many tasks in health sciences. The aim of this talk is to make an overview of some of AI techniques and approaches and their applications in medical informatics and e-health. The talk covers the following applications: (a) expert systems approach for cancer and heart diagnosis, (b) ontological engineering approach for breast cancer knowledge management, and (c) mining patient data using rough sets theory to determine thrombosis disease.

SECURITY IN CLOUD COMPUTING

Prof. Anu Gokhale

Illinois State University, United States

Abstract



Security is one of the biggest issues for cloud computing. Cloud enables users to remotely store their data and enjoy on-demand high quality applications and services from a shared pool of configurable computing resources. Cloud security and privacy concerns are arising because user data and applications are residing on providers' premises. This talk will discuss a model that separates data encryption and decryption from data storage by enabling the user to encrypt and decrypt data with the cloud provider storing encrypted data. The merits and demerits of competing security architectures and algorithms will also be addressed. The audience will come away with an understanding of current practices and the research that is currently underway to address the security challenges in cloud computing.

APPLIED VISUAL COMPUTING: CHALLENGES AND OPPORTUNITIES

Prof. Hassan Ugail

Director of Centre for Visual Computing

Abstract



In today's world, visual computing techniques that integrate 2D and 3D image processing, machine learning, 3D graphics and modelling technologies are taking centre stage in the progress of different disciplines. The focus of this talk will be on some of the recent developments in the field of visual computing particularly concentrating on the work being carried out at the Centre for Visual Computing at University of Bradford. Particularly, the focus will be on new technologies for processing large data sets in 3D, visualisation and analysis.

This talk will also focus on some of the challenges in the investigated topics and will identify open research problems that need to be resolved along with some future research directions.

Biography: Professor Hassan Ugail is the director for the Centre for Visual Computing at University of Bradford, UK. Prof. Ugail has a first class BSc Honours degree in Mathematics from King's College London and a PhD in the field of geometric design from the School of Mathematics at the University of Leeds. Prof. Ugail's research interests include geometric modelling, functional design, applications of geometric modelling to real time interactive and parametric design as well as applications of geometric modelling to general engineering problems. He has completed a number of funded projects in these areas of research and published heavily in these fields. He is also heavily involved in knowledge transfer activities and has several patents on novel techniques relating to geometry modelling, animation and 3D data exchange. He has also won the University of Bradford Vice-Chancellor's Excellence in Knowledge Transfer Award for his outstanding contribution to research and knowledge transfer activities.

WHAT IS AFTER IMAGE PROCESSING!

Dr. Mohamed Kayyali

Ongoing for Harvard University, United States

Abstract



Since the last century and founding the concept of CG, image processing becomes one of among the fastest approach in computer vision and wider after the CCD&CMOS. But what people and nations are expecting to see after a decade more-less? What is after image processing, scanning, camera, robotics and all acquisitions? What are the new concepts in relation to physics and geometry? What type of applications we can see in future after image processing?.

Workshops

WORKSHOP ON ARTIFICIAL INTELLIGENCE IN E-LEARNING AND EDUCATION

Invited Speakers

Prof. Dr. Abdel-Badeeh M. Salem

Confirmed Instructors

Symposium Scientific Committee

Artificial intelligence methodologies and techniques give e-learning systems added computing capability, allowing them to exhibit more intelligent behaviour. On the other side, the convergence of artificial intelligence, data mining, machine learning, educational technology and web science is enabling the creation of a new generation of knowledge-based tutoring and e-learning systems. The objective of AleLE'15 symposium is to bring together scientists engaged in Educational Technology, Computational Thinking, Web Technology, Knowledge Engineering and Artificial Intelligence. It will provide a forum for identifying important contributions and opportunities for recent research on the different intelligent methodologies and techniques for developing intelligent tutoring and e-Learning systems.

CONSUMER ELECTRONICS INFORMATION TECHNOLOGIES IN 21ST CENTURY COURSE OR BLESSING!

Invited Speakers

Ali Alao

We shall explore consumer electronics in the area computer hardware/software and apps i.e. laptops, tablet, mobile phones, GPS navigations, Google glasses, TV set, DVD player, smart kitchen, e.t.c. The Hardware that is mundane to consumers'- Smart cars - Blue motion engine! or low emission automobiles or green vehicle The automobiles of the future in Britain, the government pilot scheme to testing driverless cars in the UK next year. Bristol, Greenwich in south east London and Coventry and Milton Keynes (illustrated) will all host autonomous driving projects that will run for between 18 and 36 months starting from January 2015 . What are the implications and downside of these modern devices to the consumers?

Artificial Intelligence

Using temporal formalisms to support organizational creativity

Maria Mach-Król

Dept. of Business Informatics
University of Economics
Katowice, Poland
maria.mach-krol@ue.katowice.pl

Abstract— The paper is devoted to the question of supporting organizational creativity with temporal logics implemented in intelligent systems. It presents motivation for such a solution, and discusses possible formalisms to be used. The main aim of the paper is to present different application areas in the context of organizational creativity, where temporal logics could be successfully used.

Keywords— *temporal logic, temporal intelligent system, organizational creativity, creative and situational knowledge*

I. INTRODUCTION

Organizational creativity is a relatively new concept in the theory of management, which partially arose on the ground of knowledge management.

There are many definitions of organizational creativity, but it is commonly perceived as a team, dynamic activity, responding to changing features of organization's environment, a team process – see e.g. [1], [2].

The organizational creativity is therefore to be perceived in the context of organizational dynamics, because it depends on the situational changes and is composed of processes. Therefore while discussing the question of computer support for organizational creativity, the temporal aspects may not be omitted.

Such a way of formulating this problem – underlining its dynamic aspect – justifies a proposal of using an intelligent system with a temporal knowledge base, as a tool supporting creation and development of organizational creativity, which is understood as organizational asset (see e.g. [3], [4]).

By the system with a temporal knowledge base we will understand (slightly modifying the definition given in [5]) an artificial intelligence system, which explicitly performs temporal reasoning. Such a system contains not only fact base, rule base, and inference engine, but also directly addresses the question of time. For an intelligent system to be temporal, it should contain explicit time representations in its knowledge base – formalized by the means of temporal logics – and at least in the representation and reasoning layers.

The main aim of the paper is to present the application areas of organizational creativity, where temporal logics could be helpful.

The paper is organized as follows. In section 2 motivation for using temporal logic to support organizational creativity is presented. Section 3 contains some discussion on temporal representation of creative knowledge, and on different

application areas within organizational creativity, where temporal reasoning may be used. The next section discusses some proposals of temporal logics that may be successfully used to formalize organizational, creative knowledge. In section 5, the advantages of temporal formalization in the context of organizational creativity are pointed out. The last section contains summary and conclusions.

II. MOTIVATION

While discussing the use of any computer tool, one has to take into account first of all the features of the domain to be supported. This applies also to systems with a temporal knowledge base and their application in supporting organizational creativity.

Some elements that justify the use of an intelligent tool with direct time references, may be found in the definitions of organizational creativity:

[6] and [7] claim that the effects of organizational creativity encompass ideas and processes – which in our opinion should be referred to as creative knowledge. The knowledge is to be codified and stored in a knowledge base, and because it is a changing knowledge, the knowledge be should be a temporal one;

In the definition given by [8] the author points out that organizational creativity is more heuristic than algorithmic in nature (p. 33) – therefore it is not possible to use classical analytical tools, because heuristic tasks lack of algorithmic structure, they are complex and uncertain (see e.g. [9] p. 6);

[1] suggests that ideas born during creative processes (that is, the creative knowledge) must be adequate to the situation (p. 289). Therefore they have to change dynamically, because the situation of organization also constantly changes;

The changeability, dynamics, and process nature of organizational creativity, which justify its codification in a temporal knowledge base, are stressed in definitions given by [10], [11], [12], [13], [14];

[2] point out that organizational creativity must be analyzed on individual, group, and organizational levels. This justifies the use of a knowledge base: if the creativity (its effects) is to penetrate between the levels, to support collaboration, a system with a temporal knowledge base enables such penetration;

The justification for using temporal formalisms for codifying of creative knowledge may be found in the definitions given by [15], and [16], where authors point the badly structured nature of creative problems. One of temporal formalisms' advantages is the possibility to formalize unstructured problems.

While reading many authors' discussions on the essence of organizational creativity, one sees that this is primarily team activity. As it has been said above, the effect of this activity may be referred to as "creative knowledge", which itself generates new ideas, concepts, and solutions. To do so, the creative knowledge must be first codified, and next disseminated. This justifies the use of a knowledge base system. But the creative knowledge changes in time, due to several reasons.

First, organizational creativity is a process, therefore its effects are subject to change. Moreover, the process encompasses solving problems that also change, because the organization's environment changes [5] p. 13-15, [17], p. 150, 176.

Second, each knowledge – including the creative one – changes simply with the passage of time, with the flow of new information about objects [18].

Third, organizational creativity is linked with dynamics, which can be seen e.g. in the assets approach to this creativity or in the requirement of adapting creative knowledge to situational context.

The assets view of organizational knowledge and creativity the dynamics is expressed by a constant improvement of these assets to keep up with the changes in organizations and their environment – see e.g. [3]. In this way organizations adapt themselves to changes [4]. Such an adaptation occurs in time, therefore organizational creativity is connected with temporality. Moreover, assets must be developed up, therefore organizational creativity and its artifacts are dynamic.

The efforts of capturing assets' dynamics may be seen in such areas, as assets' approach, dynamic econometrics – see e.g. [19] or dynamic economics – see e.g. [20]. But these are solutions aimed only at codification and analysis of quantitative phenomena. Knowledge – including the creative one – is of qualitative nature, therefore to codify, to analyze, and to reason about it qualitative tools are needed. One of such tools is temporal logic, which enables to formalize qualitative knowledge, and also considers time. This tool is used to formalize knowledge in temporal knowledge bases. The detailed discussion on different temporal formalisms may be found e.g. in [18], [21] or [22].

All the above leads to conclusion that a knowledge base system is not enough to support organizational creativity, because classical knowledge bases do not support time. Therefore in this paper we propose the use of a temporal knowledge base system, as defined earlier. Such system is able to perform the tasks arising from the characteristics of organizational creativity and its artifacts..

III. TEMPORAL REPRESENTATION OF CREATIVE AND SITUATIONAL KNOWLEDGE

There exists an abundant literature on using temporal logics for knowledge representation and reasoning, not only in intelligent systems with a single knowledge base, but also in distributed systems, agent systems or systems coordinating robots' activities. It may be noticed a similarity between these tasks and the support of organizational creativity, because, generally speaking, it is necessary to:

- represent dynamic knowledge (about the environment),
- represent agents' beliefs (and their change),
- coordinate the functioning of elements in distributed systems.

Similar tasks are linked with supporting organizational creativity: it is necessary to represent creative knowledge and its changes, to represent knowledge about dynamic situation of organization, to coordinate activities of creative processes participants and to enable their communication.

Table I presents a survey of selected applications of temporal logics, together with their reference to organizational creativity. The references illustrate, how it would be possible to transfer concepts from the literature to the system with temporal knowledge base, supporting organizational creativity.

As it can be seen from the above, the use of temporal logics for supporting distributed, team activities is not a new idea. Temporal applications for engineering domain were present already years ago, from the very beginning of research in this area. Only short ago there came up attempts for using this formalization for management, see e.g. [5], [32] and other works by this author. In this paper the novelty lies in the application of temporal formalism – to the organizational creativity and creative knowledge.

The applications of temporal logics enumerated in Table 1 became an inspiration to elaborate a concept of using temporal logics in the intelligent system supporting organizational creativity, because this creativity is a dynamic, team process, that proceeds in the interaction with changeable, unsure organization's environment.

In the context of our proposal, the attention should be paid to the work [33], in which the authors suggest using ontology and temporal logic to model complex activities based on temporal knowledge. The process of organizational creativity is such an activity, and creative knowledge as well as situational

knowledge are both temporal. In this paper we do not address the question of creative knowledge ontology, but it should be noted that an attempt to create such an ontology and to link it

with the selected ontology of time, as well as with a selected temporal formalism, is a very interesting research problem.

TABLE I. SELECTED APPLICATIONS OF TEMPORAL LOGICS.

Author	Application	Reference to organizational creativity
[23]	The use of temporal logic for reasoning about possible behavior of distributed hybrid systems	Temporal logic as a tool for reasoning about possible development directions of organizational creativity and creative knowledge
[24]	The use of TPL (Temporal Pattern Logic) modification named FTPL for dynamic reconfiguration of system components	Dynamic reconfiguration of creative knowledge sub-bases
[25]	The use of temporal logic for conceptual modeling of data	The use of temporal logic for conceptual modeling and representation of creative knowledge
[26]	LTL – Linear Temporal Logic used to control robots in an uncertain environment	The use of temporal logic for formalizing knowledge about (uncertain) organization’s situation
[27]	Specifications in temporal logic, to control probabilistic systems operating in dynamic, partially known environment	Specifications in temporal logic to control changes of creative knowledge and knowledge about organization’s situation
[28]	The use of incremental temporal logic to control robots interacting with dynamic agents	Temporal logic as a tool for incremental representation of dynamic creative knowledge
[29]	Axioms of temporal logic used to self-control of logical agents	Temporal logic used to control changes of creative knowledge sub-bases
[30]	Approach arising from LTL to synthesize communication strategies, and control strategies in a robots’ team, depending on the environment	The use of temporal logic to drive the communication of a team in an organizational creativity process, and/or the use of temporal logic to drive the communication with a system with a temporal knowledge base
[31]	Control of dynamic systems with the use of temporal logic	Temporal logic as a tool for manipulating knowledge in an intelligent system

^a. Source: own elaboration.

IV. PROPOSALS OF TEMPORAL FORMALIZATION FOR ORGANIZATIONAL CREATIVITY

In the literature there are proposals of using very different temporal logics. In our paper we propose to use – in order to represent creative and situational knowledge – the situation calculus (and the programming language Golog, aimed at implementing programs written in situation calculus). The reasons for choosing this formalism are as follows:

- Situation calculus is a formalism for describing dynamic knowledge [34],
- Situation calculus has been successfully used to describe agents’ collaboration [35].

The most often proposals of using situation calculus concern technical domain – e.g. inference on qualitative information in order to control robots [36], beliefs change of robots [37], [38] – the second work contains also a description of application performing tasks of changing the beliefs; change of agents’ beliefs with incomplete or imprecise knowledge about the

environment [39]. These are proposals that may inspire the use of situation calculus in a temporal intelligent system supporting organizational creativity. But a special attention should be paid to the work [40], in which the authors propose to use situation calculus to support agents’ collaboration, where „agents” are teams of employees in an organization, created to exchange knowledge and intellectual assets while performing complex tasks. Therefore the situation calculus may be used in a system with temporal knowledge base, supporting the process of organizational creativity.

The situation calculus has been proposed by J. McCarthy in the sixties [41], and further developed by this author together with P. Hayes [42].

The situation calculus is a second order logic, aimed at describing dynamically changing world. Every possible world history is a path of succeeding situations. A special situation is situation denoted S_0 , so-called initial situation, in which no changes have occurred yet. Every next situation results from performing some actions. Formally speaking, $do(a, s)$ denotes a situation that occurs after performing action a in the situation s . Therefore actions cause changes in the world. Each action has

This paper has been supported by a grant: „Methodology for Computer Supported Organizational Creativity” from National Science Centre in Poland, 2013/09B/HS4/00473.

preconditions – conditions that must be fulfilled if the action is to be performed. Formally, action precondition is a sentence of the form:

$$\text{Poss}(A(x_1, \dots, x_n), s) \equiv \Pi_A(x_1, \dots, x_n, s) \quad (1)$$

Where:

A – n-ary function symbol,

Π_A – a formula uniform in s, with free variables form within x_1, \dots, x_n, s .

The causal laws are expressed using the so-called effect axioms. Another group of important axioms are the frame axioms, indispensable for establishing, which features remain unchanged independently of performing a given action. The name of these axioms comes from the commonly known problem in the temporal community – namely the frame problem – see e.g. [43], [44], [45]. As the number of frame axioms is generally infinite (because there is an infinite number of features in the world, which remain unaffected by an action), newer approaches propose to use so-called successor state axioms, which describe direct effects of actions much more precisely [34]. Formally, a successor state axiom for (n+1)-ary relational fluent F is a sentence of the form:

$$F(x_1, \dots, x_n, \text{do}(a, s)) \equiv \Phi_F(x_1, \dots, x_n, a, s) \quad (2)$$

Where $\Phi_F(x_1, \dots, x_n, a, s)$ is a formula uniform in s.

A successor state axiom for (n+1)-ary functional fluent f is a sentence of the form:

$$f(x_1, \dots, x_n, \text{do}(a, s)) = y \equiv \Phi_f(x_1, \dots, x_n, y, a, s) \quad (3)$$

where Φ_f is a formula uniform in s.

An important feature of the situation calculus in the context of supporting organizational creativity is the possibility to formulate statements concerning causality. This calculus has been used by Reiter to describe changes in a database [34], it also was the basis for many other logical solutions.

The detailed description of the situation calculus, and its formalization may be found e.g. in [34] or [35].

It is an open question whether while deploying a system with a temporal knowledge base the situation calculus will be the only sufficient formalization, or whether it will be necessary to use its extended version, namely temporal situation calculus. The original situation calculus is a so-called action language, aimed at formalizing actions and their effects, treating actions as primary causes of changes in the world. In [46] the author introduced an extension to the situation calculus, expressing time directly. He claimed that introducing the actual time line to the situation calculus, enables specification of behavior rules (p. 52), and this would make easy formalization of creative knowledge. The solution to this dilemma will be possible during the planned practical research.

V. ADVANTAGES OF TEMPORAL FORMALIZATION IN THE CONTEXT OF ORGANIZATIONAL CREATIVITY

Summing up the above discussion, the advantages of temporal formalization in an intelligent system supporting organizational creativity should be stressed.

Using temporal representation is well motivated, there are a lot of theoretic works on temporal formalisms and their features, also temporal formalisms have been used in many domains. It is certain, that temporal representation of a domain – including organizational knowledge – has many advantages. They can be divided into several groups:

- a) Basic advantages – concerning temporal representation itself, independently from where it is used; these basic advantages also are the origin of advantages from other groups;
- b) Advantages concerning representation of change;
- c) Advantages concerning representation of causal relationships.

Time, as a dimension, is a basis for reasoning about action and change – only a proper use of temporal dimension allows for representation of change and its features, as e.g. its scope or interactions caused by change [47]. Such explicit temporal reference is possible through the use of a temporal formalism, where time is a basic variable. Moreover, time may be treated in different ways, e.g. may have several different structures, which is necessary in more complex reasoning tasks, e.g. creative ones. The advantages of non-standard time representation, e.g. branching time, and its application for managerial tasks, are presented in [48], and organization of a creative process is one of managerial tasks.

Temporal logic allows encoding both qualitative and quantitative temporal information, as well as relationships among events during the creative process, therefore it is easy to express such relations, as “shorter”, “longer”, “simultaneously”, “earlier” etc. This in turn implies easiness of arranging phenomena in time, even if they overlap – Allen’s interval algebra is an example of a formalism which allows such arrangements.

Temporal formalization makes possible to encode discrete and dense changes (according to a model of time adopted), allows for describing change as a process, and for reasoning about causes, effects and directions of change, e.g. changes of creative ideas or in the creative domain itself.

As time is the fourth dimension of the world, it may not be omitted during the reasoning process; otherwise the perspective of analysis would be too narrowed. The temporal dimension allows the organizational creativity support system to “learn”: the system collects cases concerning e.g. ideas (or the creative domain) being represented, traces their evolution and thanks to this is able to generate new solutions.

It has been already said that temporal representation makes possible to represent change as a process. It is so, because with temporal logic, processes can be modeled explicitly – therefore knowledge on their temporal aspect, their interactions, on concurrent processes is easily expressed [49]. Models of processes are useful for describing dense phenomena, as for example economic ones.

Temporal logic gives us richer – temporal aspect included – formalization of domain knowledge, it also gives us “knowledge on knowledge”: combining temporal operators with formal knowledge representation one can formulate assertions about creative knowledge evolution in a system. Van Benthem presents an example of such combination, suggesting combining temporal and epistemic logic [18], p. 335. Placing creative knowledge in time treated as a basic dimension, one can add new creative knowledge to a base, not removing the “old” one, and

with no risk of inconsistencies. Temporal logic, as a knowledge representation language, should provide both explicit knowledge and access to tacit one. Temporal logic, which has reasoning rules built in, is able to provide this property.

Summing up, it should be pointed out that temporal formalisms meet the requirements of knowledge representation in artificial intelligence, such as:

- expressing imprecise and unsure knowledge,
- expressing “relations” of knowledge (e.g. A occurred before B”, that very often have no explicit dates;
- different reasoning granulations,
- modeling of persistence.

TABLE II. ADVANTAGES OF TEMPORAL FORMALIZATION IN DIFFERENT ASPECTS OF ORGANIZATIONAL CREATIVITY.

Aspect (application)	Advantages
General	explicit temporal references: time as a basic notion, formalization of alternative states during creative process, representation of changes in relations between creative features/objects, representation of qualitative and quantitative information and temporal relations in creative knowledge easy representation of overlapping events in the creative process, history of features, events, objects, relations in the creative knowledge
Knowledge base	representation of creative beliefs, explicit modeling of creative processes, persistence modeling, “knowledge about creative knowledge”, Access to tacit creative knowledge, No contradictions between old and new creative knowledge
Representation of changes	Discrete and continuous changes in the creative domain, Description of changes as a process, Reasoning about reasons, effects and directions of changes in the creative domain
Representation of causal relationships	Easy description of causal relationships in the creative domain Retro- and proactive events, “if-then” analysis of creative ideas
Reasoning	“learning” of the organizational creativity support system, Tracing creative features’ evolution, Reasoning about dynamic aspects of creative phenomena, Reasoning about sequences of events in the organizational creativity support system, The notion of “possibility”, Qualitative reasoning about creative domain, Simulation of human commonsense and creative reasoning

^b Source: Own elaboration.

The above postulates are met e.g. by Allen’s interval algebra [49]. Therefore enriching an organizational creativity support system with temporal formalisms would allow for taking into account the temporal dimension of creative knowledge, its changes and evolution/development. In this way the creative knowledge, and organizational creativity processes may be managed more effectively. Advantages of temporal formalization in different areas of organizational creativity are presented in Table II.

The research conducted by [50] lets us formulate the postulates for a temporal logic, used to represent creative and situational knowledge. Taking into account the anthropocentric approach proposed by Kalczynski and Chou, one may state the following:

- a) It is not important whether the formalism comes from 1st order predicate logic or from modal logic,
- b) Formalisms based on change (e.g. situation calculus and its mutations) are closer to human reasoning about time, than formalisms based on time,

- c) Basic temporal entities are not important – using time points or intervals or both to represent creative domain will depend on particular needs – e.g. people during the creative process do not analyze economic texts in the context of temporal entities,
- d) Representation of knowledge should be based on qualitative or mixed approaches; the quantitative approach may be considered a special case of the qualitative one – every numeric feature can be represented in a qualitative manner.

Summing up, while choosing the temporal logic to represent creative and situational knowledge, one should first of all consider the way humans perceive temporal aspects of the world. The possibilities of particular temporal logics are less important. In this way the knowledge representation – close to human perception – will enable a more understandable temporal reasoning in the intelligent system.

VI. CONCLUDING REMARKS

The main aim of this paper was to discuss the possibility of supporting organizational creativity with temporal reasoning. To do this, one should consider a system with temporal knowledge base, formalized in a selected temporal logic or logics.

The use of temporal formalization is justified, as organizational creativity is a dynamic process, moreover, the “product” of this process – namely creative knowledge – is also dynamic. The creative knowledge is of qualitative nature, therefore using temporal logics seems natural, as these formalisms are dedicated and elaborated to represent qualitative phenomena, and their change in time.

We have pointed out several application domains within the context of organizational creativity, where temporal logics may be used. We also proposed to choose the situation calculus or its temporal extension. Of course this is not the only possible choice. As organizational creativity is strictly connected with commonsense reasoning, one may consider also choosing the event calculus [51]. The choice of a formalism to be implemented in a temporal intelligent system will be the subject of future research.

The main research in the future will concern a conceptual model of a temporal intelligent system for organizational creativity support. We also plan to encode some portion of a creative knowledge in a temporal formalism, and check whether temporal reasoning on such encoded knowledge is possible.

REFERENCES

- [1] K. L. Unsworth, „Unpacking Creativity,” *Academy of Management Review*, Vol. 26, No. 2, pp. 286-297, 2001.
- [2] C. Andriopoulos and P. Dawson, *Managing Change, Creativity and Innovation*. Second Edition, Los Angeles/London/New Delhi/Singapore/Washington DC: SAGE Publications, 2014.
- [3] R. Krupski, ed., *Rozwój szkoły zasobowej zarządzania strategicznego*, Wałbrzych: Wałbrz. Wyz. Szk. Zarz. i Przedsięb., 2011.
- [4] D. G. Sirmon, M. A. Hitt, R. D. Ireland and B. A. Gilbert, „Resource Orchestration to Create Competitive Advantage: Breadth, Depth, and Life Cycle Effects,” *Journal of Management*, Vol. 37, No. 5, pp. 1390-1412, September 2011.
- [5] M. A. Mach, *Temporalna analiza otoczenia przedsiębiorstwa. Techniki i narzędzia inteligentne*, Wrocław: Wydawnictwo AE, 2007.
- [6] R. W. Woodman, J. E. Sawyer and R. W. Griffin, „Toward a Theory of Organizational Creativity,” *The Academy of Management Review*, Vol. 18, No. 2, pp. 293-321, April 1993.
- [7] C. E. Shalley, L. L. Gilson and T. C. Blum, „Matching Creativity Requirements and the Work Environment: Effects on Satisfaction and Intentions to Leave,” *Academy of Management Journal*, Vol. 43, No. 2, pp. 215-223, 1 April 2000.
- [8] T. M. Amabile, *Creativity in Context: Update to The Social Psychology of Creativity*, Boulder: Westview Press, 1996.
- [9] A. Aggarwal, „A Taxonomy of Sequential Decision Support Systems,” In: *Proc. IS-2001: 4th Annual Informing Science Conference*, Kraków, 2001.
- [10] E. C. Martins and F. Terblanche, „Building organisational culture that stimulates creativity and innovation,” *European Journal of Innovation Management*, Vol. 6, No. 1, pp. 64 - 74, February 2003.
- [11] L. D. Alvarado, „The creative organizations as living systems,” In: *Understanding and evaluating creativity*, S. Torre and V. Violant, eds., Malaga, Ediciones Algiba, 2006, pp. 375-382.
- [12] G. Hirst, D. v. Knippenberg and J. Zhou, „A Cross-Level Perspective on Employee Creativity: Goal Orientation, Team Learning Behavior, and Individual Creativity,” *Academy of Management Journal*, Vol. 52, No. 2, pp. 280-293, 1 April 2009.
- [13] M. Baer, „Putting Creativity to Work: The Implementation of Creative Ideas in Organizations,” *Academy of Management Journal*, Vol. 55, No. 1, pp. 1102-1119, 1 October 2012.
- [14] M. Basadur, T. Basadur and G. Licina, „Organizational Development,” In: *Handbook of Organizational Creativity*, M. D. Mumford, ed., London/Waltham/San Diego, Elsevier Inc., 2012, pp. 667-703.
- [15] M. D. Mumford, I. C. Robledo and K. S. Hester, „Creativity, Innovation and Leadership: Models and Findings,” In: *The SAGE Handbook of Leadership*, A. Bryman, D. Collinson, K. Grint, B. Jackson and M. Uhl-Bien, eds., London, SAGE Publications Ltd., 2011, pp. 405-421.
- [16] M. D. Mumford, K. E. Medeiros and P. J. Partlow, „Creative Thinking: Processes, Strategies, and Knowledge,” *The Journal of Creative Behavior*, Vol. 46, No. 1, pp. 30-47, March 2012.
- [17] S. Czaja, *Czas w ekonomii. Sposoby interpretacji czasu w teorii ekonomii i w praktyce gospodarczej*, Wrocław: Wydawnictwo Uniwersytetu Ekonomicznego, 2011.
- [18] J. van Benthem, „Temporal Logic,” In: *Handbook of Logic in Artificial Intelligence and Logic Programming. Volume 4: Epistemic and Temporal Reasoning*, D. M. Gabbay, C. J. Hogger and J. A. Robinson, eds., Oxford, Clarendon Press, 1995, pp. 241-350.
- [19] J. Jakubczyk, *Wprowadzenie do ekonometrii dynamicznej*, Warszawa-Wrocław: Wydawnictwo Naukowe PWN, 1996.
- [20] J. Adda and R. Cooper, *Dynamic Economics. Quantitative Methods and Applications*, Cambridge, Mass.: The MIT Press, 2003.
- [21] R. Klimek, *Wprowadzenie do logiki temporalnej*, Kraków: Uczelniane Wydawnictwa Naukowo-Dydaktyczne AGH, 1999.

- [22] M. Fisher, D. Gabbay and L. Vila, eds., *Handbook of Temporal Reasoning in Artificial Intelligence*, Amsterdam-Boston-Heidelberg: Elsevier B.V., 2005.
- [23] P. Hou and H. Zheng, „Quantified Differential Temporal Dynamic Logic for Verifying Properties of Distributed Hybrid Systems,” In: *Logical Foundations of Computer Science*, LNCS Vol. 7734, S. Artemov and A. Nerode, eds., Berlin-Heidelberg, Springer, 2013, pp. 234-251.
- [24] J. Dormoy, O. Kouchnarenko and A. Lanoix, „Using Temporal Logic for Dynamic Reconfiguration of Components,” In: *7th International Conference on Formal Aspects of Component Software*, Berlin-Heidelberg, 2012.
- [25] A. Artale, R. Kontchakov, F. Wolter and M. Zakharyashev, „Temporal Description Logic for Ontology-Based Data Access,” In: *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, Beijing, 2013.
- [26] S. C. Livingston, R. M. Murray and J. Burdick, „Backtracking temporal logic synthesis for uncertain environments,” In: *Proc. ICRA-12: International Conference on Robotics and Automation*, Saint Paul, MN, 2012.
- [27] T. Wongpiromsarn and E. Frazzoli, „Control of Probabilistic Systems under Dynamic, Partially Known Environments with Temporal Logic Specifications,” In: *51st Annual Conference on Decision and Control*, Maui, HI, 2012.
- [28] T. Wongpiromsarn, A. Ulusoy, C. Belta, E. Frazzoli and D. Rus, „Incremental Temporal Logic Synthesis of Control Policies for Robots Interacting with Dynamic Agents,” In: *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012.
- [29] S. Constantini, „Self-checking Logical Agents (Extended Abstract),” In: *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems*, Saint Paul, MI, 2013.
- [30] Y. Chen, X. Ding, A. Stefanescu and C. Belta, „A Formal Approach to Deployment of Robotic Teams in an Urban-Like Environment,” In: *Distributed Autonomous Robotic Systems*, Springer Tracts in Advanced Robotics 83, A. Martinoli, F. Mondada, N. Correll, G. Mermoud, M. Egerstedt, M. Hsieh, L. Parker and K. Stoy, eds., Berlin Heidelberg, Springer, 2013, pp. 313-327.
- [31] E. Wolff, U. Topcu and R. Murray, „Robust Control of Uncertain Markov Decision Processes with Temporal Logic Specifications,” In: *IEEE Conference on Decision and Control*, Grand Wailea, Maui, Hawaii, 2012.
- [32] M. Mach-Król, „Prospects of Using Temporal Logics for Knowledge Management,” In: *Advances in Business ICT*, M. Mach-Król and T. Pelech-Pilichowski, eds., Heidelberg, Springer, 2014, pp. 41-52.
- [33] G. Okeoy, L. Chen, H. Wang and R. Sterritt, „A Hybrid Ontological and Temporal Approach for Composite Activity Modelling,” In: *11th International Conference on Trust, Security, and Privacy in Computing and Communications*, Liverpool, 2012.
- [34] R. Reiter, *Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems*, Cambridge, MA: MIT Press, 2001.
- [35] J. Claßen, „Planning and Verification in the Agent Language Golog,” Aachen University: doctoral dissertation, Aachen, 2013.
- [36] S. Schiffer, A. Ferrein and G. Lakemeyer, „Reasoning with Qualitative Positional Information for Domestic Domains in the Situation Calculus,” *Journal of Intelligent and Robotic Systems*, Vol. 66, No. 1-2, pp. 273-300, April 2012.
- [37] V. Belle and H. Levesque, „Reasoning about Continuous Uncertainty in the Situation Calculus,” In: *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, Beijing, China, 2013.
- [38] M. Pagnucco, D. Rajaratnam, H. Strass and M. Thielscher, „Implementing Belief Change in the Situation Calculus and an Application,” In: *Logic Programming and Nonmonotonic Reasoning*, P. Cabalar and T. Son, eds., Berlin Heidelberg, Springer LNCS 8148, 2013, pp. 439-451.
- [39] J. Delgrande and H. Levesque, „Belief Revision with Sensing and Fallible Actions,” In: *Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, 2012.
- [40] A. Toniolo, T. J. Norman, K. Sycara and J. H. Farrington, „Agent Support for Collaboration in Complex Deliberative Dialogues.(Extended Abstract-PhD Thesis),” IFAAMAS.org, 2012.
- [41] J. McCarthy, „Situations, actions and causal laws,” In: *Semantic Information Processing*, M. Minsky, ed., Cambridge, Mass., MIT Press, 1968, pp. 410-417.
- [42] J. McCarthy and P. Hayes, „Some philosophical problems from the standpoint of artificial intelligence,” *Machine Intelligence*, Vol. 4, pp. 463-502, 1969.
- [43] P. Hayes, *The Frame Problem and Related Problems on Artificial Intelligence*, Stanford: Stanford University, 1971.
- [44] Y. Xu and P. Wang, „The frame problem, the relevance problem, and a package solution to both,” *Synthese*, Vol. 187, No. 1, pp. 43-72, 2012.
- [45] A. Zambak, „The Frame Problem,” In: *Philosophy and Theory of Artificial Intelligence*, V. Müller, ed., Berlin Heidelberg, Springer, 2013, pp. 307-319.
- [46] J. Pinto, „Temporal reasoning in the situation calculus,” University of Toronto: doctoral dissertation, Toronto, 1994.
- [47] L. Vila, „Formal Theories of Time and Temporal Incidence,” In: *Handbook of Temporal Reasoning in Artificial Intelligence*, M. Fisher, D. Gabbay and L. Vila, eds., Amsterdam, Elsevier, 2005, pp. 1-24.
- [48] M. Mach-Król, „Nonlinear Time Ontology for Economic Reality Description,” In: *The 6th International Conference on Information Technology*, Amman, 2013.
- [49] J. Allen, „Towards a General Theory of Action and Time,” *Artificial Intelligence*, Vol. 23, No. 2, 1984.
- [50] P. Kalczynski and A. Chou, „Temporal Document Retrieval Model for Business News Archives,” *Information Processing & Management*, Vol. 41, No. 3, pp. 635-650, 2005.
- [51] E. T. Mueller, *Commonsense Reasoning*, San Francisco: Morgan Kaufmann/Elsevier, 2006.

Multi Objects Tracking in Nighttime Traffic Scenes

Mohamed Taha, Hala H. Zayed

Computer Science Dept.
Faculty of Computers & Informatics, Benha University
{mohamed.taha, hala.zayed}@fci.bu.edu.eg

Taymoor Nazmy and M. E. Khalifa

Computer Science Dept., Basic Science Dept.
Faculty of Computer & Information Sciences, Ain Shams University
ntaymoor19600@gmail.com, esskhalifa@cis.asu.edu.eg

Abstract—As road networks become more congested, traffic surveillance using computer vision techniques is increasingly important. Traffic surveillance can help in improving road network efficiency, re-routing traffic when accidents occur and minimizing delays. Although, there are many algorithms developed to detect and track moving vehicles in daytime, only a handful of techniques have been proposed for nighttime traffic scenes. In the night environment, the moving vehicles are commonly identified by detecting and locating vehicle headlights and taillights. This paper proposes an effective method for detecting and tracking moving vehicles in nighttime. The proposed method identifies vehicles by detecting and locating vehicle lights using automatic thresholding and connected components extraction. Detected lamps are then paired using rule based component analysis approach and tracked using Kalman Filter (KF). The automatic thresholding approach provides a robust and adaptable detection process that operates well under various nighttime illumination conditions. Furthermore, most nighttime tracking algorithms detects vehicles by locating either headlights or rear lights. However, the proposed method has the ability to track vehicles through detecting vehicle headlights and/or rear lights. Several experiments are presented that demonstrate the feasibility and the effectiveness of the proposed method to detect and track vehicles in various nighttime environments.

Keywords—Traffic Surveillance; Nighttime Surveillance; Vehicles Tracking; Vehicles Detection; Nighttime Tracking; Multi Objects Tracking

I. INTRODUCTION

Computer vision techniques have been widely used in many applications to automatically characterize the environment and understand the scene. Intelligent transportation systems, traffic surveillance, driver assistance systems, autonomous vehicle guidance, and road traffic information systems have recently received significant attention from the computer vision community. All of these applications need some kind of information about moving vehicles. Traffic data are critical for traffic management and other transportation applications. In the past decades, loop detectors or supersonic wave detectors were used to estimate the traffic flow or traffic density on a road [1]. However, these methods are limited to the number of vehicles passing through the detection regions and are difficult to apply for vehicle classification, vehicle speed detection, and vehicle motion analysis [2]. Today, digital cameras are the most popular traffic sensors used for collecting traffic data. They have the ability to capture not only traffic volumes, but also speeds, vehicle classifications, queue lengths, control delays, and other traffic parameters. These parameters can be obtained through detecting and tracking vehicles using different computer vision techniques. Video camera based systems are now smarter, highly advanced, and yet more comprehensive than ever before. The information embedded in the video frames allows identifying and classifying the vehicles effectively. The

temporal continuity among video frames can help in enhancing the accuracy during vehicle detection process [3].

Of all computer vision techniques, object tracking has been very active research in the last years. Despite being classic computer vision problem, tracking is largely unsolved. There are still many challenges that need more research including illumination effects (such as shadows, changes in ambient lighting), scene clutter (such as objects in background, other moving scene objects), changes in target appearance (such as the addition or removal of clothing, and changing facial expressions), occlusions, and simultaneously tracking multiple targets with similar appearance.

Although visual surveillance is a very active topic in computer vision, it primarily focuses on algorithms designed for daytime [4, 5]. Nighttime vehicle surveillance is still important because high traffic flows as well as incidents can happen during night on city roads or highways. In addition, under bad-illuminated condition in the nighttime road environment, the obvious features of vehicles which are effective for detecting in daytime become invalid in nighttime road environment. Most recent studies on vehicle detection adopt frame differencing, and background subtraction techniques to extract the features of moving vehicles from traffic scenes. Although these techniques are effective for vehicle detection in daytime, they become inefficient in nighttime illumination conditions. This is because the

background scenes are greatly affected by the varying lighting effect of moving vehicles [4]. In other words, the daytime traffic surveillance systems exploit the greyscale, color and motion information to detect and analyze the vehicles. Nevertheless, under the nighttime traffic environment, this information become meaningless where the camera images have very low contrast and a weak light sensitivity. In these conditions, the vehicles can only be identified by locating their headlights and rear lights. These are the only visual features of the vehicles at night and under darkly illuminated conditions. Furthermore, there are strong reflections on the roads surface, which complicate the problem. The moving reflections of the headlights can introduce a lot of foreground or background ambiguities. Therefore, vehicle nighttime detection and tracking is still an open research area with many potential for improvement.

Two different types of nighttime environment can be found when considering vehicles detection and tracking: highway and urban road. Each type has its own characteristics that affect the detection process. The highway environment is an unlit scene where there are no street lights and the only features visible are the headlights and their reflections. The headlights appear as a bright round blob in contrast to the dark surroundings. On the other hand, the urban road is a lit scene where the streets are illuminated by the public light poles like most of the urban areas. In this type of scenes, the background, pedestrians and other objects are also visible. Here the headlights are not clearly visible, especially when the vehicles are also in white or some light colors. Hence the complexity of extracting the headlights from the images is increased. Fig. 1 show a typical examples of nighttime traffic scene for a highway and an urban road [4]. In the urban environment, many similar light blob features could be mistakenly detected as vehicle.

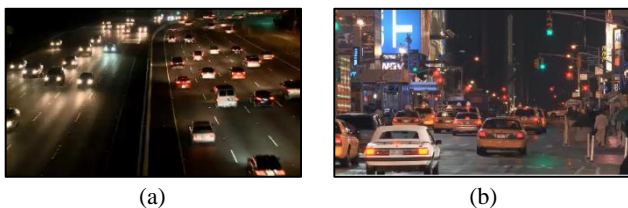


Fig 1. Typical examples of nighttime traffic scenes. (a) A highway. (b) An urban road.

Here, we primarily focus on detecting vehicles lights because they have high intensity during nighttime. These lights can correctly be discriminated from reflection on the road surface. This paper proposes an effective method for detecting and tracking moving vehicles in nighttime. The proposed method identifies vehicles by detecting and locating vehicle lights using automatic thresholding and connected components extraction. Detected lamps are then paired using rule based component analysis approach and tracked using Kalman Filter (KF). Fixed thresholds usually limit the performance of most existing vehicle detection algorithms in extracting and pairing vehicle lights. This retards the algorithms from being adapted to real traffic scenes. However, the automatic thresholding approach employed by the proposed method provides a robust

and adaptable detection process that operates well under various nighttime illumination conditions. Moreover, most nighttime tracking algorithms detects vehicles by locating either headlights or rear lights. However, the proposed method has the ability to track vehicles through detecting vehicle headlights and/or rear lights. This means that it works for both oncoming vehicles (headlights detection) and preceding vehicles (rear lights detection).

For clarity of presentation, the paper is organized as follows: Section II explores the related work found in the literature concerning vehicle detection in nighttime. Section III presents the proposed method in detail. Section IV discusses the experimental results and the performance evaluation of the proposed method.

II. RELATED WORK

Recently, research in the area of vehicle tracking both at night and during daytime has grown rapidly. There is a rapid need for such systems in many applications such as driver assistance systems and surveillance systems. Performance indexes required by these systems include high recognition rates, real-time implementation, robustness for variant environments, and feasibility under poor visibility conditions. Although lot of published work has been done on vehicles detection and tracking in daytime, there are very less number of researchers who worked on nighttime scenarios. In this section, an overview of the state-of-the-art methods is given for on-road nighttime vehicle detection and tracking. In fact, there is no general method for solving this problem although some patterns can be observed. Many works have been put forward in the literature for nighttime traffic surveillance. Actually, vehicle lights have been widely used as discernment features for nighttime vehicle detection applications in traffic monitoring systems and driver assistance systems [6-14]. Most of these methods use morphological operations to extract candidate headlight objects and then perform shape analysis, template matching, or pattern classification to find the paired headlights of moving vehicles. Nevertheless, there are many problems due to complex real-time conditions. Therefore, vehicle nighttime detection and tracking is still an open area with many potential for improvement.

Salvi [8] presents a traffic surveillance system for detecting and tracking moving vehicles in various nighttime environments. The algorithm is composed of four steps: headlight segmentation and detection, headlight pairing, vehicle tracking, and vehicle counting and detection. First, a fast segmentation process based on an adaptive threshold is applied to extract bright objects of interest. The extracted bright objects are then processed by a spatial clustering and tracking procedure that locates and analyzes the spatial and temporal features of vehicle light patterns, and identifies and classifies moving cars and motorbikes in traffic scenes. However, the classification function of the algorithm needs to be improved to enhance the classification capability on different vehicle types, such as buses, trucks, and light and heavy motorbikes

Wang et al. [9] propose a region tracking-based vehicle detection algorithm during nighttime via image processing techniques. Their algorithm is based on detecting vehicle taillights and use it as the typical feature. The algorithm uses the existing global detection algorithm to detect and pair the taillights. When the vehicle is detected, a time series analysis model is introduced to predict vehicle positions and the possible region (PR) of the vehicle in the next frame. Then, the vehicle is only detected in the PR.

Zhang et al. [10] propose a nighttime traffic surveillance system, which consists of headlight detection, headlight tracking and pairing, camera calibration and vehicle speed estimation. First, a vehicle headlight is detected using a reflection intensity map and a reflection suppressed map based on the analysis of the light attenuation model. Second, the headlight is tracked and paired by utilizing a bidirectional reasoning algorithm. Finally, the trajectories of the vehicle's headlight are employed to calibrate the surveillance camera and estimate the vehicle's speed. The disadvantage of this system is when one headlight of the vehicle is occluded by other vehicles, it cannot be paired with other headlights.

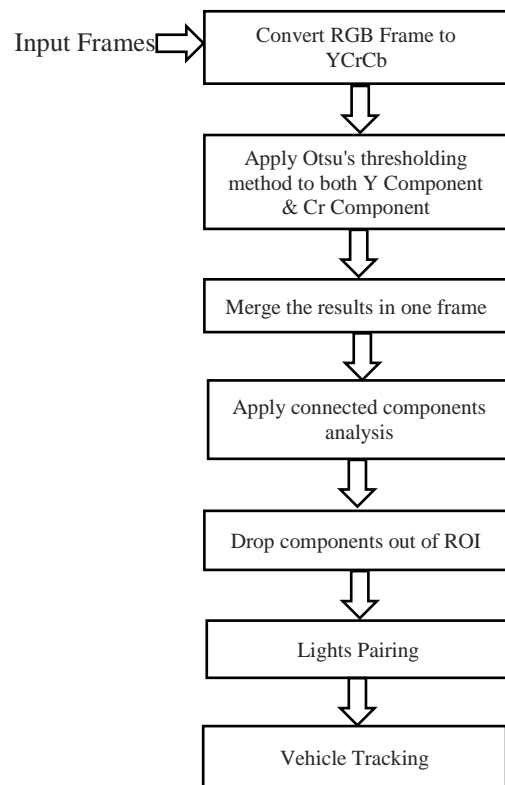
Chen et al [11] present a traffic surveillance system for detecting and tracking moving vehicles in nighttime traffic scenes. Their method identifies vehicles by detecting and locating vehicle headlights and taillights using image segmentation and pattern analysis techniques. First, a fast bright-object segmentation process based on automatic multilevel histogram thresholding is applied to effectively extract bright objects of interest. The extracted bright objects are then processed by a spatial clustering and tracking procedure that locates and analyzes the spatial and temporal features of vehicle light patterns, and identifies and classifies moving cars and motorbikes in traffic scenes. The disadvantage of this system is it can identify only cars and motorbikes. It fails to detect and track other vehicles.

O'Malley et al [12] present a system to detect and track vehicle rear-lamp pairs in forward-facing color video. A standard low-cost camera with a complementary metal-oxide semiconductor (CMOS) sensor and Bayer Red-Green-Blue (RGB) color filter is used for full-color image display or other color image processing applications. Rear-facing lamps are segmented from low-exposure forward-facing color video using a red-color threshold. Lamps are paired using color cross-correlation symmetry analysis and tracked using Kalman filtering. A tracking-based detection stage is introduced to improve robustness and to deal with distortions caused by other light sources and perspective distortion, which are common in automotive environments. The drawback of this system is that it fails to detect target vehicles that were greater than 50 m away due to lack of intensity or insufficient resolution of vehicles

III. PROPOSED METHOD

The basic idea of the proposed method is based on the fact that vehicle lights (headlights and rear lights) are strong and consistent features that can be used to reveal the presence of a moving vehicle at night. Vehicle lights appear as the brightest

regions, whether on highways or on urban roads. Regardless of the type of street lighting or the weather conditions, the vehicle



lights features remain relatively stable. Fig. 2 shows the block diagram of the proposed method.

Fig. 2. The block diagram of the proposed method

In order to detect the vehicle lights, it is common for image processing techniques to use some form of thresholding. However, the RGB color space is not ideal for the task of color thresholding. It is difficult to set and manipulate color parameters due to high correlation between the Red, Green, and Blue channels. Hence, the proposed method starts with converting the color space of the video frame from RGB to YCbCr. Y is the luminance component while Cb and Cr are the Blue-difference and Red-difference Chroma components. The main advantage of converting the frame to YCbCr color space is that this color space is characterized by its ability to separate the light intensity component from the chrominance. The Y component gives all information about the brightness, while the Cb (Blue) and Cr (Red) components are independent from the luminosity influence. However, in the RGB color space, each component (Red, Green and Blue) has a different brightness.

In nighttime traffic, vehicle lights appear as the brightest pixels in the video frames. Headlight objects are bright and therefore appear white in color while the core part of rear light object is red. In order to detect these pixels, the proposed method uses Otsu's thresholding technique [15, 16] to both Y component and Cr component. Vehicle headlights can be detected by applying the Otsu's technique to the luminance component Y while the rear lights (red light sources) can be

detected by applying the Otsu's technique to the Red-difference Chroma channel Cr. The thresholding is performed using the following equation:

$$g(x, y) = \begin{cases} 1 & f(x, y) > T \\ 0 & \text{otherwise} \end{cases}$$

Where $f(x, y)$ denotes the intensity of the pixel (x, y) in the video frame, $g(x, y)$ indicates the corresponding segmentation result after thresholding and T is the threshold value. To make the thresholding process more robust, the threshold value T should be automatically selected with each frame. The manual threshold setting method and offline learning based method cannot adapt to the variation of the environment in real-time. So in the proposed algorithm, a dynamic threshold is calculated using Otsu's method [15, 16]. It is designed to select the optimum threshold for separation into two classes based upon maximizing the variance between them. It involves iterating through all the possible threshold values and calculating a measure of spread (intra-class variance) for the pixel levels on each side of the threshold, i.e. the pixels that fall either in foreground or in background. The aim of this step in the algorithm is to find the threshold value where the sum of foreground and background spreads is at its minimum. It does not depend on modelling the probability density functions; however, it assumes a bimodal (i.e., two classes) distribution. Fig. 3 shows an illustration of the thresholding process. A video frame of a highway nighttime traffic surveillance is shown in Fig. 3.a. The results of applying Otsu's thresholding technique to both Y component and Cr component are shown in Fig. 3.b and 3.c respectively. In order to extract all the bright pixels (headlights and rear lights) found in the frame, the thresholding results of both the luminance(Y) and the red component (Cr) are combined to form a single frame as shown in Fig. 3.d and 3.e.

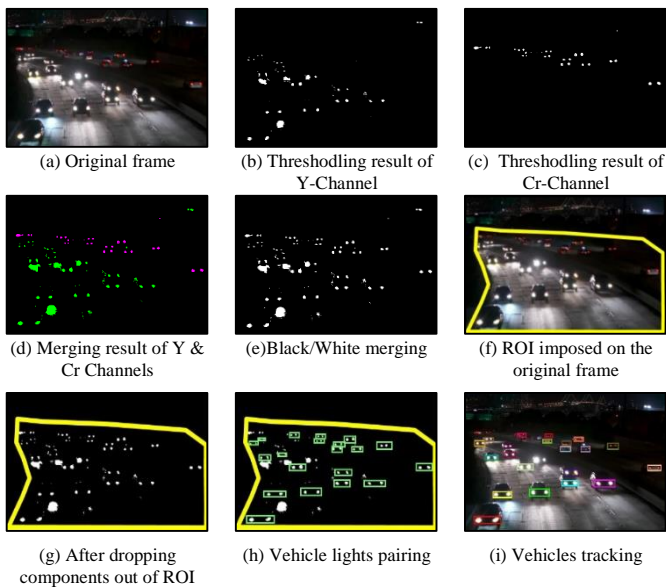


Fig. 3. Moving vehicles detection of a highway frame

Fig. 4 shows a similar illustration of the thresholding process but on a video frame of an urban road. By investigating the two

Fig. 3 and 4, it should be noted that some interferential objects, such as street lamps and traffic lights, are also detected at the top side of the frame especially in the urban road (Fig. 4.e). Utilizing the thresholding method extracts all the bright pixels. Hence, further filtering is required as there are many potential light sources that are not vehicle lamps. To filter out these objects, the proposed method applies two consecutive steps: connected component analysis [17, 18] and Region Of Interest (ROI) filtering. First, a connected component extraction process is performed to locate the connected components of the bright objects. Extracting these components clarifies the significant features of location, dimension, shape, and pixel distribution associated with each connected component. Second, the ROI filtering is applied to each video frame to exclude any connected component with a location out of the detection region (see Fig. 3.f, 3.g, 4.f, 4.g). The detection region should cover the lanes that are being monitored. It is usually set at the lower part of the image. It can be predetermined either manually during the setup of the surveillance camera or automatically by using lane detection algorithms [19-23]. Then, the detection is only performed in the ROI. Hence, after masking outside the ROI, the scene becomes simpler, since out-of-ROI distracting objects, such as street lamps, are removed. The ROI not only can reduce complexity in searching for vehicle candidates but also can decrease the false positive detection rate. At the same time, ROI definition speeds up the processing time as only a part of original image is processed.

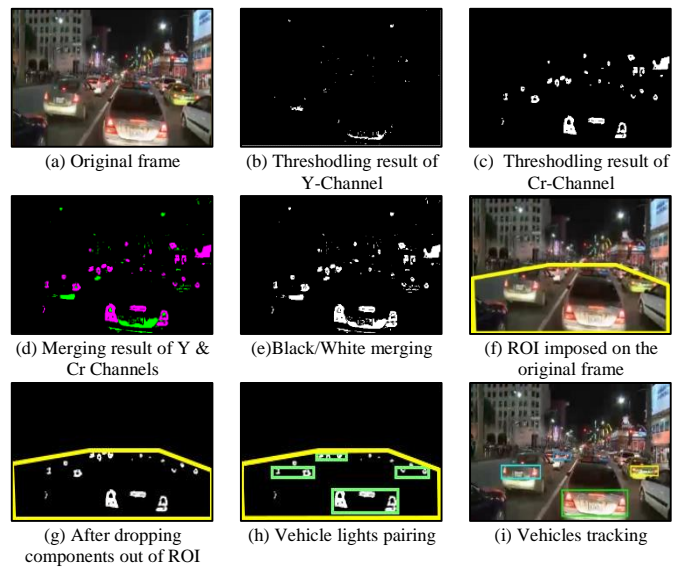


Fig. 4. Moving vehicles detection of an urban road frame

The next step of the proposed method is to pair the identified vehicle lights in order to start tracking. The proposed method adopts the rule based component analysis approach [23-25] where the identified vehicle lights can be paired with each other if certain rules are satisfied. The pairs of vehicle lights must have some common properties. Hence, two connected components are said to belong to the same vehicle if the following rules are satisfied [9, 24, 25]:

- The components must be horizontally close to each other and the vertical and horizontal positions should be considered.
- The components are of similar size.
- The width to height ratio of the bounding box enclosing the two components must be greater.
- Area of the pixels must be similar.
- The symmetry condition must be satisfied.

Fig. 3.h and 4.h shows the results of vehicle lights pairing step. Finally, the proposed method uses Kalman Filter (KF) to perform the tracking process. Vehicles are tracked using the four parameters of a bounding box surrounding the lamp pair (x-position, y-position, width, and height). Kalman filter [26], also known as linear quadratic estimation (LQE), is an algorithm that uses a series of measurements observed over time, containing noise (random variations) and other inaccuracies, and produces estimates of unknown variables that tend to be more precise than those based on a single measurement alone [27-29]. Kalman filter provides a general solution to the recursive minimized mean square linear estimation problem [30]. The mean square error will be minimized as long as the target dynamics and the measurement noise are accurately modelled. Kalman filter is composed of two steps [31]: prediction and correction. In the prediction step, the location of an object being tracked is predicted in the next frame while in the correction step, the object in the next frame within designated region is identified. A set of KFs is used to keep track of a variable and unknown number of moving targets [27]. Each time a new observation is received, it is associated to the correct track among the set of the existing tracks or if it represents a new target, a new track has to be created. The tracking results are shown in Fig. 3.i and 4.i. Moreover, it should be mentioned that several advantages are gained for using Kalman filter [28-29]. First, prediction using the basic Kalman filter is extremely fast and requires little memory. This makes it a convenient form for online real time processing. Second, it is easy to formulate and implement given a basic understanding. Third, an error estimate is associated with each prediction. Fourth, these predictions can be computed recursively, bounding the time and memory needed for computation.

IV. EXPERIMENTAL RESULTS

In order to analyze the performance of the proposed algorithm, several experiments were conducted to evaluate the nighttime vehicle detection and tracking performance achieved by the proposed method. The experiments were implemented on a 2.27GHz Intel Core i5 PC with 4GB memory, running under Windows 8 Enterprise. The algorithm is coded using MATLAB 8.1.0.604 (R2013a).

Establishing standard test beds is a fundamental requirement to compare algorithms. Unfortunately, there is no standard dataset to compare the results and efficiency of nighttime vehicles detection and tracking algorithms. This was a major difficulty in order to compare the performance of the

proposed method with its counterparts. Most of algorithms found in the literature record their videos by their own. To evaluate the performance of the proposed method, two sets of experiments are conducted. The first set of experiments are performed over a self-collected and prepared dataset. It consists of 14 video clips containing both urban and highway scenes (downloaded from <http://www.videoblocks.com/>). All the video sequences are with a frame rate equal to 30 frames per second and the size of the grabbed image sequence is 480 × 270 pixels with 24-bit true color. The video clips are selected with different traffic density (high – medium- low). The ground truth for each video clip of the dataset was prepared manually. A detailed description of the dataset is found in Table I. The second set of experiments are performed over the testing video data used in [9]. It consists of four video sequences captured in an urban roadway environment. Two videos of them (video a and video b) contain only one moving vehicle in the scene while the others (video c and video d) contain from two to four moving vehicles in the scene. The frame rate of each video is 30 frames per second and the size of the grabbed image sequence is 720 × 480 pixels with 32-bit true color.

TABLE I. DETAILED DESCRIPTION OF THE DATASET

No	Video sequence name	No of frames	No of vehicles	Vehicles move direction	Scene type
1	Above LA Highway Traffic	379	23	Unidirectional	Highway
2	Highway LA Overpass	332	53	Bidirectional	Highway
3	LA Highway Bend Traffic	404	96	Bidirectional	Highway
4	LA Highway Bend	394	64	Bidirectional	Highway
5	Slow Moving 101 North Traffic	530	41	Bidirectional	Highway
6	Slow Night Commute In Cali	311	74	Bidirectional	Highway
7	Cars On LA Highway	653	112	Bidirectional	Highway
8	Night Time Traffic on Snowy Downtown Street in Homer	812	4	Bidirectional	Urban
9	Slow Moving Los Angeles Traffic	586	35	Unidirectional	Urban
10	Nighttime Traffic in Aspen	557	7	Unidirectional	Urban
11	Roadway Traffic at Night in Snowy Small Town	599	3	Bidirectional	Urban
12	Seattle Airport Control Tower and Traffic at Night	752	5	Bidirectional	Urban
13	Taxi Cabs and Traffic in Times Square	490	9	Unidirectional	Urban
14	Traffic on Busy Times Square Street	531	10	Unidirectional	Urban

Fig. 5 and Fig. 6 show the results of applying the proposed tracking method to both highway road and urban road traffic scenes of our dataset respectively. The first row shows the original frame of a nighttime surveillance video. The second row display the results after applying the first five steps of the

proposed method: applying Otsu's thresholding to both Y Component & Cr Component, merging the results in one frame, applying connected components analysis, and excluding components out of ROI. The third row shows the results of pairing detected vehicle lights where the green rectangles indicate the vehicle lights that have been paired. The fourth row shows the tracking results. The experimental results demonstrate that the proposed method can robustly detect and track vehicles in different nighttime traffic environments. Note that the video sequences have different illumination conditions. Hence, the desired thresholds to detect vehicle lights should be different for each video frame. Using the adaptive Otsu's thresholding technique in the proposed method, the desired threshold suitable for each frame can be found efficiently without any manual intervention.

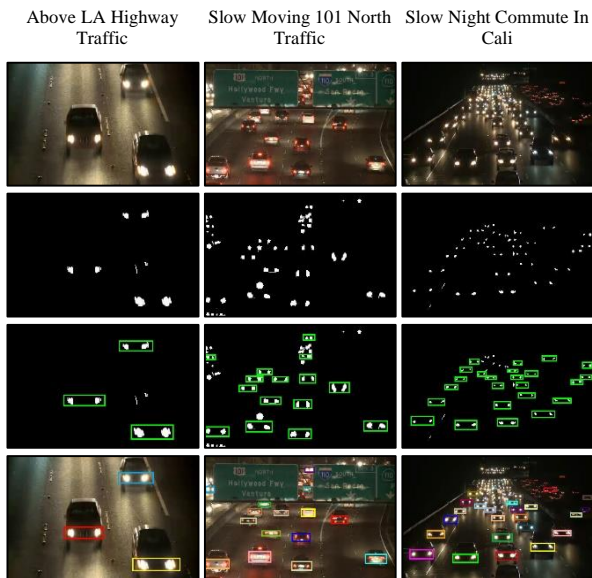


Fig. 5. Moving vehicles tracking in highway road traffic scenes of the dataset

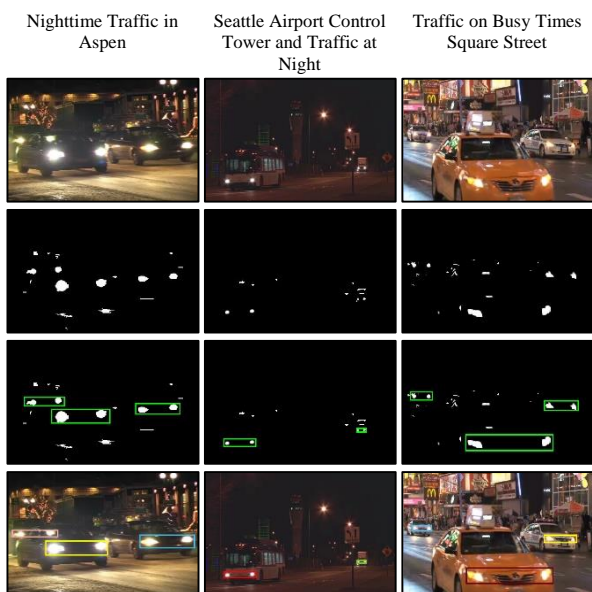


Fig. 6. Moving vehicles tracking in urban road traffic scenes of the dataset

Table II shows the quantitative results of the proposed method for vehicle tracking in different nighttime traffic environments. The average tracking rates of the proposed method are 96.27% and 95.76% for both urban and highway scenes respectively in our dataset. Almost all the vehicle lights can be detected and the false tracking of vehicles occurs when the vehicles move side by side or when there exist some moving reflection objects on the road. This in turn may cause some false pairing. However, the effect of interferential objects such as street lamps are attenuated by the step of excluding all detected components outside the ROI.

TABLE II. TRACKING RATE OF THE PROPOSED METHOD FOR OUR DATASET

No	Video sequence name	No of vehicles	No of correctly tracked vehicles	Tracking rate (%)
1	Above LA Highway Traffic	23	20	86.96%
2	Highway LA Overpass	53	52	98.11%
3	LA Highway Bend Traffic	96	94	97.92%
4	LA Highway Bend	64	63	98.44%
5	Slow Moving 101 North Traffic	41	39	95.12%
6	Slow Night Commute In Cali	74	74	100.00%
7	Cars On LA Highway	112	109	97.32%
8	Night Time Traffic on Snowy Downtown Street in Homer	4	4	100.00%
9	Slow Moving Los Angeles Traffic	35	32	91.43%
10	Nighttime Traffic in Aspen	7	7	100.00%
11	Roadway Traffic at Night in Snowy Small Town	3	3	100.00%
12	Seattle Airport Control Tower and Traffic at Night	5	5	100.00%
13	Taxi Cabs and Traffic in Times Square	9	8	88.89%
14	Traffic on Busy Times Square Street	10	9	90.00%

The following part evaluates the performance of the proposed method and compares it to the region tracking-based vehicle detection algorithm presented by Wang et al. [9]. Fig. 7 shows the comparative results of nighttime vehicle tracking for running the two methods on the test sequences used in [9]. The first column of the figure shows the original frame. The second column shows the results of applying the region tracking-based vehicle detection algorithm. The third column shows the results of applying the proposed method. As it can be noted from the figure, the proposed methods successfully detects and tracks all vehicles appeared in the scene. However, the region tracking-based algorithm does not perform well in detecting all vehicles under some complicated nighttime traffic scenes, and some vehicles are missed. This is because the proposed method applies the adaptive thresholding step to both Y Component and Cr Component of the video frame and merge the two results in one frame. However, the region tracking-based vehicle detection algorithm applies the thresholding on the gray scale image of the video frame. Gray scale image based segmentation succeeds in segmenting white pixels but fails in segmenting red pixels especially in low

illumination conditions. Pixels with high red color component and low green and blue color components are bright red in color but their corresponding gray scale values can be low. Hence, these pixels face a difficulty to be detected.

Table III shows the tracking rates achieved when running both the Region Tracking-Based Vehicle Detection Algorithm [9] and the proposed method on the testing sequences used in [9].

As the table indicates, both algorithms succeed in tracking all the vehicles appeared in the video a and video b because both videos contain only one vehicle to be tracked. However, when the number of vehicles increases, the performance of the Region Tracking-Based Vehicle Detection Algorithm is degraded while the proposed method still successfully detects and tracks almost all vehicles.

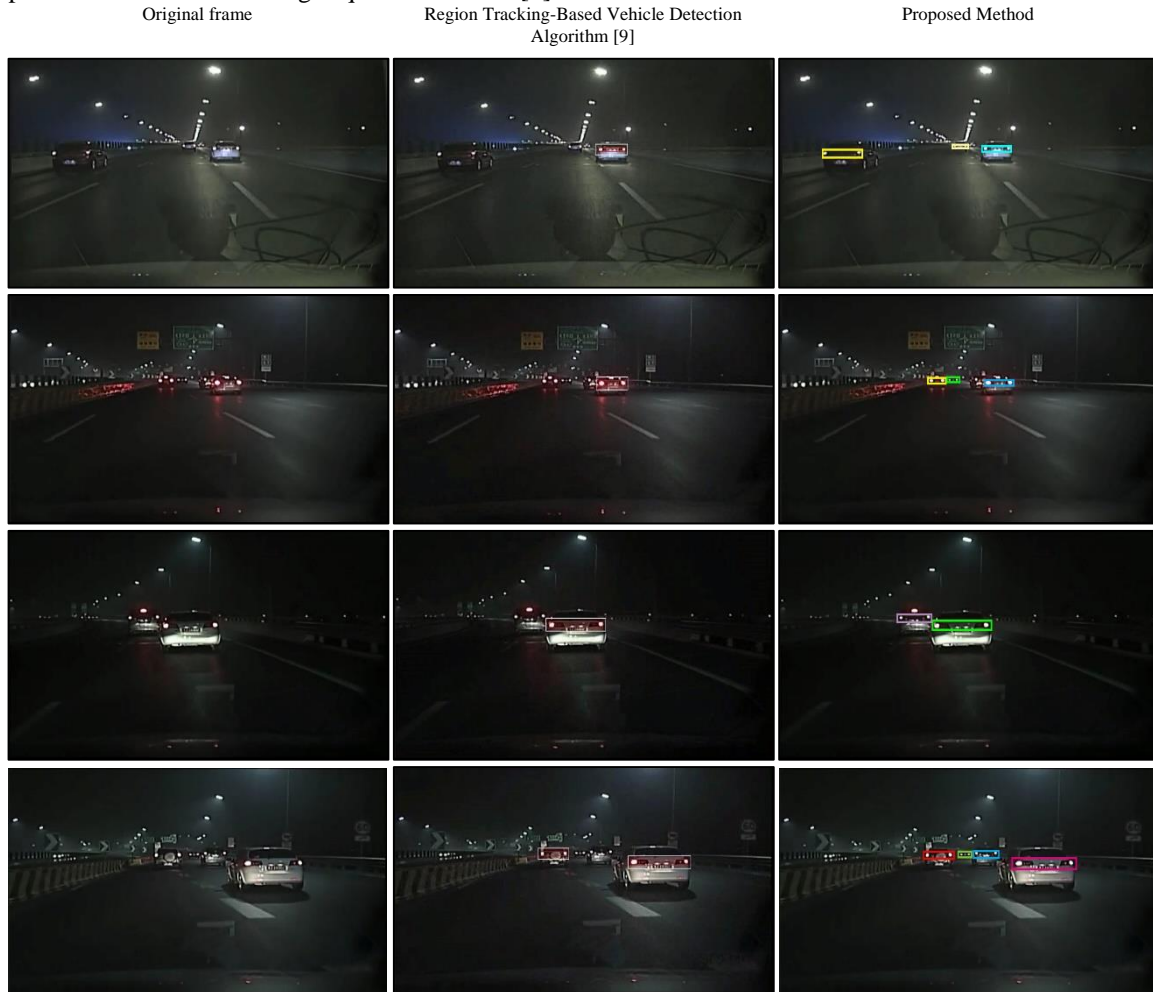


Fig. 7. Comparative results of vehicle detection and tracking in nighttime traffic scenes on test sequences used in [9]

TABLE III. TRACKING RATE OF THE PROPOSED METHOD FOR TEST SEQUENCES USED IN [9]

No	Video sequence name	Tracking rate (%)	
		Region Tracking-Based Vehicle Detection Algorithm [9]	Proposed method
1	Video a	100%	100%
2	Video b	100%	100%
3	Video c	98.67%	99.1%
4	Video d	93.45%	97.23%

Based on the above results, in most cases, the proposed method can detect and track vehicles correctly. However, it may fail in some cases. First, it cannot be used to detect the parked vehicles or vehicles with low visibility lights. This is because parked vehicles usually have lights switched off and

the proposed method mainly depends on detecting vehicle lights in order to detect and track vehicles. However, this does not pose any problem to the performance of the proposed method because tracking is concerned only with moving objects. Second, when an object occludes one lamp of the vehicle, the other lamp can still be correctly detected but cannot be paired with other vehicles lights in the scene. Finally, some vehicles may have four headlights, and these four headlights may be paired as two vehicles. To solve this defect, the vehicle's length information can be incorporated in the pairing process. After the surveillance camera is calibrated, the distance between two pairs of headlights can be determined. Therefore, the four headlights can be paired as one vehicle if the distance is less than some threshold.

V. CONCLUSION

In this paper, a method for detecting and tracking moving vehicles in nighttime is proposed. The proposed method is able to detect and track the vehicles in low visibility conditions at nighttime. It recognizes vehicles by detecting vehicle lights using automatic thresholding and connected components extraction. Next, it finds pairs of vehicles lights to estimate vehicle locations using rule based component analysis and it employs Kalman Filter (KF) in the tracking process. The automatic thresholding approach provides a robust and adaptable detection process that operates well under various nighttime illumination conditions. Moreover, most nighttime tracking algorithms detects vehicles by locating either headlights or rear lights. Nevertheless, the proposed method has the ability to track vehicles through detecting vehicle headlights and/or rear lights. Experimental results demonstrate that the proposed method is feasible and effective for vehicle detection and identification in various nighttime environments.

ACKNOWLEDGMENT

We would like to express our sincere appreciation to Dr. Jianqiang Wang at Tsinghua University and Dr. Xiaoyan Sun at Suzhou INVO Automotive Electronics Co., for their help in providing their dataset to be used in testing the performance of the proposed method.

REFERENCES

- [1] Padmavathi, S, Keerthana Gunasekaran, "Night Time Vehicle Detection for Real Time Traffic Monitoring Systems: A Review" In the International Journal of Computer Technology & Applications, Volume 5, Number 2, PP. 451-456, April 2014.
- [2] Yen-Lin Chen, Bing-Fei Wu, Hao-Yu Huang, and Chung-Jui Fan, "A Real-Time Vision System for Nighttime Vehicle Detection and Traffic Surveillance," In Proceedings of IEEE Transactions on Industrial Electronics, Volume 58, Issue 5, pp.2030-2044, May 2011.
- [3] Sayanan Sivaraman and Mohan M. Trivedi, "A Review of Recent Developments in Vision-Based Vehicle Detection," IEEE Intelligent Vehicles Symposium, June 2013.
- [4] Kovacic, Kristian, Ivanjko, Edouard and Gold, Hrvoje. "Computer Vision Systems in Road Vehicles: A Review," In the Proceedings of the Croatian Computer Vision Workshop, October 2013.
- [5] Wherever Z. Sun, G.Bebis, and R.Miller, "On-road vehicle detection: A review," In Proceedings of IEEE Transactions Pattern Analysis and Machine Intelligence, volume 28, Number 5, pp. 694 -711, May 2006.
- [6] Weihong Wang, Chunhua Shen, Jian Zhang, S. Paisitkriangkrai, "A Two-Layer Nighttime Vehicle Detector," In Proceedings of the International Conference of Digital Image Computing: Techniques and Applications (DICTA '09), Melbourne, Australia, pp. 162 - 167, December 2009.
- [7] K. Robert, "Night-Time Traffic Surveillance a Robust Framework for Multivehicle Detection, Classification and Tracking," In Proceedings of the sixth IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS '09), Genova, Italy, pp. 1-6, September 2009.
- [8] G. Salvi, "An Automated Nighttime Vehicle Counting and Detection System for Traffic Surveillance," In Proceedings of the 2014 International Conference on Computational Science and Computational Intelligence (CSCI '14), Volume 01 , Volume 01 , March 2014
- [9] Jianqiang Wang, Xiaoyan Sun, Junbin Guo, "A Region Tracking Based Vehicle Detection Algorithm in Nighttime Traffic Scenes", In the International Journal of Sensors, Volume 13, Issue 12, pp. 16474-16493, December 2013.
- [10] Wei Zhang, Q. M. J. Wu, Guanghui Wang, and Xinge Yu, "Tracking and Pairing Vehicle Headlight in Night Scenes," In Proceedings of IEEE Transactions on Intelligent Transportation Systems, Volume 13, Issue 1, March 2012.
- [11] Yen-Lin Chen, Bing-Fei Wu, Hao-Yu Huang, and Chung-Jui Fan, "A Real-Time Vision System for Nighttime Vehicle Detection and Traffic Surveillance," In Proceedings of IEEE Transactions on Industrial Electronics, Volume 58, Issue 5, pp.2030-2044, May 2011.
- [12] Ronan O'Malley, Edward Jones and Martin Glavin, "Rear-Lamp Vehicle Detection and Tracking in Low - Exposure Color Video for Night Conditions," In Proceedings of IEEE Transactions on Intelligent Transportation Systems, Volume 11, Issue 2, pp.453-462, June 2010.
- [13] Chun-Che Wang, Shih-Shinh Huang, Li-Chen Fu, "Driver Assistance System for Lane Detection and Vehicle Recognition with Night Vision," In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2005), Alberta, Canada, pp. 3530-3535, August 2005.
- [14] Yen-Lin Chen, Yuan-Hsin Chen, Chao-Jung Chen, Bing-Fei Wu, "Nighttime Vehicle Detection for Driver Assistance and Autonomous Vehicles," In Proceedings of 18th International Conference on Pattern Recognition (ICPR 2006), Hong Kong, Volume 1, pp. 687-690, August 2006.
- [15] S. Kumar, M. Pant, A. Ray, "Differential Evolution Embedded Otsu's Method for Optimized Image Thresholding," In Proceedings of the 2011 World Congress on Information and Communication Technologies (WICT), Mumbai, India, pp. 325 - 329, December 2011.
- [16] Mehmet Sezgin, and Bulent Sankur, "Survey Over Image Thresholding Techniques and Quantitative Performance Evaluation," In the Journal of Electronic Imaging, Volume 13, Issue 1, pp. 146-168, January 2004.
- [17] Lifeng Hea, Yuyan Chaob, Kenji Suzukic, and Kesheng Wud, "Fast Connected-Component Labeling," In the International Journal of Pattern Recognition, Volume 42, Issue 9, pp. 1977-1987, September 2009.
- [18] Kenji Suzukia, Isao Horibaa, and Noboru Sugieb, "Linear-Time Connected-Component Labeling Based on Sequential Local Operations," In the International Journal of Computer Vision and Image Understanding, Volume 89, Issue 1, pp. 1-23, January 2003.
- [19] Aharon Bar Hillel, Ronen Lerner, Dan Levi, Guy Raz, "Recent Progress in Road and Lane Detection: a Survey," In Machine Vision and Applications Journal, Volume 25, Issue 3, pp. 727-745, April 2014.
- [20] Mohamed Hammami, Nadra Ben Romdhane, Hanene Ben-Abdallah, "An Improved Lane Detection and Tracking Method for Lane Departure Warning Systems," In the International Journal of Computer Vision and Image Processing, Volume 3, Issue 3, pp. 1-15, July 2013.
- [21] Sibel Yenikaya, Gökhan Yenikaya, Ekrem Düven, "Keeping the Vehicle on the Road: A Survey on On-Road Lane Detection Systems," In the International Journal of ACM Computing Surveys (CSUR), Volume 46, Issue 1, Article No. 2, October 2013.
- [22] Jianyu Yang, Zhuo Li, Liangchao Li, "Lane Detection Based on Classification of Lane Geometrical Model," In Proceedings of the IEEE 11th International Conference on Signal Processing (ICSP), Beijing, China, pp. 842 - 846, October 2012.
- [23] H. Deusch, J. Wiest, S. Reuter, M. Szczot, M. Konrad, K. Dietmayer, "A Random Finite Set Approach to Multiple Lane Detection," In Proceedings of the 15th International IEEE Conference on Intelligent Transportation Systems (ITSC), Anchorage, Alaska, USA, pp. 270-275, September 2012.
- [24] Shifu Zhou, Jianxiong Li, Zhenqian Shen and Liu Ying "A Night time Application for a Real-Time Vehicle Detection Algorithm Based on Computer Vision", Research Journal of Applied Sciences, Engineering and Technology, Volume 5, Number 10, pp. 3037-3043, March 2013.
- [25] Yen-Lin Chen, "Nighttime Vehicle Light Detection on a Moving Vehicle using Image Segmentation and Analysis Techniques", WSEAS Transactions On Computers, Volume 8, Issue 3, March 2009.
- [26] Rudolph Emil Kalman, "A New Approach to Linear Filtering and Prediction Problems," In Transactions of the ASME-Journal of Basic Engineering, Volume 82(Series D), pp. 35-45, 1960.

- [27] R. Sathya Bharathi, "Video Object Tracking Mechanism," In the Journal of Computer Engineering (IOSR-JCE), Volume 16, Issue 3, PP 20-26, May 2014.
- [28] Gregory F. Welch, "History: The Use of the Kalman Filter for Human Motion Tracking in Virtual Reality," In the International Journal of Teleoperators and Virtual Environments, Volume 18, Issue 1, pp. 72-91, February 2009.
- [29] Vahid Fathabadi, Mehdi Shabbazian, Karim Salahshour, Lotfollah Jargani, "Comparison of Adaptive Kalman Filter Methods in State Estimation of a Nonlinear System Using Asynchronous Measurements," In Proceedings of the World Congress on Engineering and Computer Science(WCECS 2009), Vol II, San Francisco, USA, October 2009.
- [30] Shih-Ku Weng, Chung-Ming Kuo, Shu-Kang Tu, "Video Object Tracking Using Adaptive Kalman Filter," In the International Journal of Visual Communication and Image Representation, Volume 17, Issue 6, pp. 1190-1208, December 2006.
- [31] Amir Salarpour, Arezoo Salarpour, Mahmoud Fathi, and MirHossein Dezfoulian, "Vehicle Tracking Using Kalman Filter and Features," In the International Journal of Signal & Image Processing (SIPIJ), Volume 2, Number 2, June 2011.

New Selection Schemes for Particle Swarm Optimization

Mohammad Mohammad Shehab

School of Computer Sciences
Universiti Sains Malaysia
Penang, Malaysia

Mohammed Azmi Al-Betar

Department of Information Technology
Al-Huson University College, Al-Balqa Applied University, Al-Huson
Irbid, Jordan

Ahamad Tajudin Khader

School of Computer Sciences
Universiti Sains Malaysia
Penang, Malaysia

Abstract— In Evolutionary Algorithms (EA), the selection scheme is a pivotal component, where it relies on the fitness value of individuals to apply the Darwinian principle of survival of the fittest. In Particle Swarm Optimization (PSO) there is only one place employed the idea of selection scheme in global best operator in which the components of best solution have been selected in the process of deriving the search and used them in generation the upcoming solutions. However, this selection process might be affecting the diversity aspect of PSO since the search infer into the best solution rather than the whole search. In this paper, new selection schemes which replace the global best selection schemes are investigated, comprising fitness-proportional, tournament, linear rank and exponential rank. The proposed selection schemes are individually altered and incorporated in the process of PSO and each adoption is realized as a new PSO variation. The performance of the proposed PSO variations is evaluated. The experimental results using benchmark functions show that the selection schemes directly affect the performance of PSO algorithm. Finally, a parameter sensitivity analysis of the new PSO variations is analyzed.

Key words— Particle Swarm Optimization; Evolutionary Algorithm; Selection Schemes; Global-best.

I. INTRODUCTION

Swarm intelligence is a discipline that deals with natural systems composed of many individuals that exhibit collective behavior, decentralized control, and self-organization [1]. Its principle depends on the method of communication and interaction between the individuals and their environment. The most important application on swarm intelligence is Particle Swarm Optimization (PSO)[2]. PSO was developed by Eberhart and Kennedy [3]. It simulates the social behavior of bird flocking or fish schooling. It is a stochastic optimization technique and is remarkably developing [4]. Its simplicity and effectiveness have caught the attention of scientists from all over the world [5]. It is used to obtain the best solution among the particles in a swarm. This solution is called the global best fitness, and the candidate solution that achieves this fitness is called the global best position [6-8]. During the improvement loop, other solutions are attracted by the global best position, whose diffusion is degraded. Thus, global selection is conducted solely from the best solution among all the solutions

(particles) to improve the next generation [9, 10]. The other solutions are ignored; therefore, the diversity of exploration may be affected, given that the search is concerned with only a single point. In other words, the global best concept of the PSO algorithm uses the search space capacity of PSO, which may be loose and therefore result in premature convergence and quick stagnation without generating efficient results.

In this study, the global best concept of PSO is substituted with a new selection scheme borrowed from Genetic Algorithm (GA). These schemes are fitness-proportional, tournament, linear rank, and exponential rank. Each scheme constructs a new PSO variation. The PSO variations are evaluated using standard mathematical optimization functions. The results show the effectiveness of the proposed selection schemes. The remainder of this study is organized as follows: Section II presents the PSO algorithm. Section III discusses the proposed selection schemes incorporated with PSO. Section IV presents the computation results, analysis, and discussion.

Section V concludes the study and provides possible directions for future study.

II. PARTICLE SWARM OPTIMIZATION ALGORITHM PRINCIPLES

The Particle Swarm Optimization PSO is a population-based optimization method proposed by Kennedy and Eberhart [11]. The behavior of PSO can be conceivable by comparing it to school of fish searching for optimal food sources, where the direction in which a fish moves is influenced by its current movement, the best food source it ever experienced.

PSO iteratively improves the accuracy of the solution to the optimization problem. Basically, optimization procedures are shown in flow chart as shown in Figure 1. These steps are described as follows:

- Initialization: The n position vectors are randomly initialized $\{X_k^{(0)}, k = 1, 2, \dots, n\}$. The elements of X_k are uniformly distributed in a suitable range. Subsequently, the n velocity vectors $\{V_k^{(0)}, k = 1, 2, \dots, n\}$ are randomly initialized with the elements uniformly distributed between the minimum and maximum values. The fitness of the particle is determined by the objective function [12]. The local best of each particle is initialized to its initial position and the global best to the best fitness among the best locals.

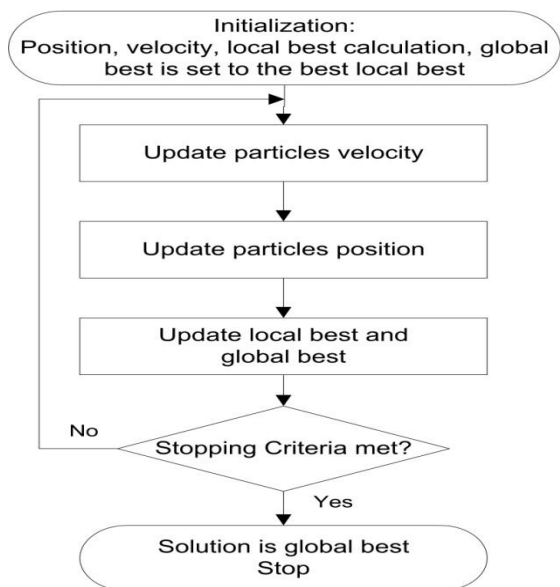


Fig. 1. Flowchart of the PSO Algorithm

- Update Velocity: Equation (1) updates the velocity of the particle [13]:

$$V_{i(t+1)} = W \cdot V_{i(t)} + C_1 \cdot r_1 (P_{ibest} - X_i) + C_2 \cdot r_2 (P_{gbest} - X_i) \quad (1)$$

where C_1 and C_2 represent the weights of the stochastic acceleration terms to the P_{ibest} and P_{gbest}

positions, r_1 and r_2 represent the random function between 0 and 1, X represents the current position of the particle, P_{ibest} is the best position of individual i until iteration t , P_{gbest} is the global best position among the whole particles, and W is the inertia weight that controls the acceleration of the particle in its optimal direction.

- Update Position: The particle position of each particle is updated depending on the updated velocity in the following equation [14]. The updated position is based on equation (2).

$$X_{i(t+1)} = X_{i(t)} + V_{i(t+1)} \quad (2)$$

- Update the Local and Global Best: The fitness of each particle is evaluated based on the new updated position. If the updated position leads to a better objective function value, the local and the global best are updated.
- Stopping Criteria: The three previous steps are repeated until the number of iterations is reached.

III. SELECTION SCHEMES

The evolutionary algorithm (EA) is generally characterized by several features, such as a high level of population diversity. It is considered to be the best method of searching [1]. EA is discriminated in diversity population to circumvent premature convergence. A higher rate of selection from an accumulative search may lead to the loss of diversity, which in turn results in premature convergence. By contrast, if the rate of selection from the existing search is small, the search depends on randomness. Thus, the slow convergence problem may be achieved.

Any selection scheme of EA consists of two main phases [15]: the selection phase, in which the selection probability is assigned to each solution in the population depending on its fitness, and the sampling phase, in which the probability controls of the sample are selected in the solutions to the next population.

Selection schemes are classified into static and dynamic schemes [16]. The selection probability of each solution in the static selection scheme is determined in advance and then remains constant during the search. Examples of this scheme include tournament selection, linear rank, and exponential rank. By contrast, the dynamic selection scheme updates the selection probability of each solution in the population at each evolution. Another classification categorizes the selection schemes into fitness-proportionate and rank-based schemes. The fitness-proportionate class calculates the selection probability based on the absolute fitness value of each individual, whereas the selection probability in the rank-based

class is determined based on fitness ranking rather than absolute fitness. A simple example of the fitness-proportionate scheme is the traditional proportional selection scheme. Some selection schemes have scaling problems that lead to premature convergence (e.g., proportionate selection). Other selection schemes suffer from the non-balance between fitness and the ability of reproduction (e.g., linear rank).

This study focuses on modifying the PSO algorithm by amending the method of selecting the global best solution. Figure 1 shows the procedure of selecting the global best. The minimum value of the global best remains stable until the end of the iterations. In this study, each proposed selection scheme is replaced with the original selection scheme. The following subsections present some of the random selection schemes that are suggested to the method used in PSO, which can be achieved by determining the working principle for each selection method when searching for sampling and selection probability for each variable in a new PSO.

A. Proportional Selection Scheme

The proportional (or roulette wheel) selection scheme proposed by Holland and John is the most traditional selection method [17]. In this method, the selection probability depends on the absolute fitness value of any solution compared with those of the other solutions in the population. The selection probability P_i for the solution i is proportional to its fitness value, which is calculated using equation (3). In Algorithm (1), r randomly picks a value uniformly from $U(0, 1)$; sum_prob has accumulative selection probabilities, where the $sum\ prob = \sum_{i=1}^j P_i$ is the accumulative selection probability of solution X^j .

$$P_i = \frac{f(x^i)}{\sum_{j=1}^{swarmsize} f(x^j)} \quad (3)$$

Algorithm 1. Pseudocode for the Proportional Selection Scheme

- 1: **Set** $r \sim U(0, 1)$.
- 2: **Set** $found = false$.
- 3: **Set** $sum_prob = 0$.
- 4: **Set** $K = 0$.
- 5: **While** ($i \leq swarm_size$) **and not** ($found$) **do**
- 6: $sum_prob = sum_prob + P_i$
- 7: **If** ($sum_prob \geq r$) **then**
- 8: $K = i$
- 9: $found = True$
- 10: **End If**
- 11: $i = i + 1$
- 12: **End While**

B. Tournament Selection Scheme

Tournament selection is among the most popular selection methods in genetic algorithms. It was initially proposed by Grefenstette and Baker [18]. Algorithm (2) shows the principle of tournament selection work, which starts from the random selection of t individuals from $P_{(t)}$ population and then proceeds to the selection of the best individual from tournament t . This procedure is repeated n times. The best choice is frequently between two individuals, and this scheme is called binary tournament, where the choice is between t individuals called tournament size [19].

Algorithm 2. Pseudocode for the Tournament Selection Scheme

- 1: **Choose** K (the tournament size) individuals from the population at random.
- 2: **Choose** the best individual from pool / tournament with probability P .
- 3: **Choose** the second best individual with probability $P^*(1-P)$.
- 4: **Choose** the third best individual with probability $P^*((1-P) \wedge 2)$.
- 5: **And so on...**

C. Linear Ranking Selection Scheme

Linear ranking is another selection scheme that was developed to overcome the disadvantages of the proportional selection scheme [20]. Rank selection schemes are developed to determine the selection probability of the solutions stored in PSO based on the solution fitness rank as shown in equation (4). The linear ranking selection scheme is based on the rank of individuals rather than on their fitness. Rank n is assigned to the best individual, whereas rank 1 is assigned to the worst individual. Thus, based on its rank, each individual i has the probability of being selected given by the expression [21].

$$P_i = \frac{rank(i)}{n * (n - 1)} \quad (4)$$

Once all individuals of the current population are ranked, the procedure of the linear rank selection scheme can be implemented based on Algorithm (3).

D. Exponential Ranking Selection Scheme

Exponential ranking selection sorts the probabilities of the ranked individuals by exponentially weighted as shown in



equation (5). The main of the exponent C is situated between 0 and 1. If $C = 1$, the difference in the selection probability between the best and the worst solutions is lost. If $C = 0$, the difference in the selection probability becomes increasingly large and follows an exponential curve along the ranked solution.

Algorithm 3. Pseudocode for Linear Ranking Selection Scheme

```

1: Set  $S_0 = 0$ 
2: For  $i=1$  to  $swarm\_size$  do
3:    $S_i = S_{i-1} + P_i$ 
4: End For
5: For  $i=1$  to  $swarm\_size$  do
6:   Generate a random number  $r \in [0, swarm\_size]$ 
7:   For each  $1 \leq j \leq swarm\_size$  do
8:     If ( $P_j \leq r$ ) do
9:       Select the  $j^{th}$  individual
10:    End If
11:   End For.
12: End For.
```

$$P_i = \frac{C^{Rank_i}}{\sum_{j=1}^{swarmsize} C^{Rank_j}} \quad (5)$$

Algorithm (4) for the exponential ranking, it is similar to that for the linear ranking. The only difference is in the calculation of the selection probabilities as stated in algorithm (5).

IV. COMPUTATIONAL RESULTS, ANALYSIS AND DISCUSSION

This section experimentally evaluates the new selection schemes. The five variations of the PSO algorithm proposed in this study are distinguished. Each variation uses a particular selection scheme that is incorporated with the PSO algorithm:

- 1) Global best Particle Swarm Optimization (GPSO): It uses the PSO algorithm with the global best selection scheme.
- 2) Proportional Particle Swarm Optimization (PPSO): It uses the PSO algorithm with the proportional selection scheme.
- 3) Tournament Particle Swarm Optimization (TPSO): It uses the PSO algorithm with the tournament selection scheme.
- 4) Linear rank Particle Swarm Optimization (LPSO): It uses the PSO algorithm with the linear rank selection scheme.

- 5) Exponential rank Particle Swarm Optimization (EPSO): It uses the PSO algorithm with the exponential rank selection scheme.

All the experiments are conducted using a computer with processor Intel(R) Core (TM) 2 Quad CPU Q9400@2.66 GHz with 4 GB of RAM and 32-bit for Microsoft Windows 7 Professional. The source code is implemented using MATLAB (R2010a). This study applies 14 benchmarks minimization problems to compare the different selection schemes using a

Algorithm 4. Pseudocode for Exponential Ranking selection scheme

```

1: Set  $S_0 = 0$ 
2: For  $i=1$  to  $swarm\_size$  do
3:    $S_i = S_{i-1} + P_i$ 
4: End For
5: For  $i=1$  to  $swarm\_size$  do
6:   Generate a random number  $r \in C$ 
7:   For each  $1 \leq j \leq swarm\_size$  do
8:     If ( $P_j \leq r$ ) do
9:       Select the  $j^{th}$  individual
10:    End If
11:   End For.
12: End For.
13: End For.
```

large test set that involves function optimization [22]. The results of the benchmark minimization functions are used to compare the default selection schemes of PSO with the proposed selection schemes in this research.

The common parameters among all the algorithms used in the experiments are set depending on the experiential instruction. The flow of the different parameter settings used to evaluate the PSO with different selection schemes is investigated. An intensive parameter analysis is conducted with various values of D , *population size*, C_1 , C_2 , and W for each PSO variation as follows: dimension size $D = (10, 20, \text{ and } 30)$ [23], *population size* = (30, 50, and 80) [24], acceleration coefficient C_1 and $C_2 = (1.5, 2, \text{ and } 2.5)$ [25], and weight $W = (0.5, 0.7, \text{ and } 0.9)$ [26]. Each run is iterated 100,000 times.

The best parameter setting for each variation is recorded in Table I. A series of experiments is then conducted using five convergence scenarios, each of which varies in terms of parameter settings, as shown in Table I. Each convergence scenario investigates the capability of the five parameters, and each of these parameters includes a set of values. For example the first scenario contains the GPSO with its best value of each parameter, as shown $D=30$, $C1=1.5$, $C2=2$, $W=0.7$ and *pop. size*= 50, These values determined as a best value for the experiments, and so on for all scenarios.

A big size of dimensions requires more function evaluations. Meanwhile, increasing the computing efforts for convergence increases the reliability of the algorithm. The main point of this study is to maintain a balance between

reliability and cost. Thus, the best value for the dimension size should be between 10 and 30 and should not be larger than 30 when the problem is complicated. The results obtained in this study are consistent with those of other researchers [4, 10, 27].

The acceleration coefficients C_1 and C_2 often have the same value. Based on the different empirical studies, the best value for the acceleration coefficients is $C_1 = C_2 = C/2$, where C is the total of acceleration coefficients. If C is small, then the algorithm explores slowly. [28] Kennedy suggests the same previous equation to determine C_1 and C_2 : ($C_1 + C_2 \leq 4.0$).

The inertia weight causes a significant increase in the convergence speed and a better balance between the exploitation and exploration of the solution space, while the complexity of the algorithm increases only slightly. Therefore, the recommended inertia weight is between 0.5 and 0.9 [26].

Selecting a population size (number of particles) of 50 is recommended for higher dimensional problems, and a population size of [30, 50] is appropriate for lower dimensional problems. The values of population size are compatible with previous studies [24, 29, 30].

TABLE I. PSO PARAMETERS SCENARIO

Scenario No.	Selection Schemes	Parameters				
		D	C_1	C_2	W	Pop. Size
Sen1	GPSO	30	1.5	2	0.7	50
Sen 2	PPSO	30	2	2	0.7	50
Sen 3	TPSO	30	2	1.5	0.7	30
Sen 4	LPSO	30	1.5	1.5	0.7	30
Sen 5	EPSO	30	1.5	2	0.7	50

A summary of the 14 global minimization benchmark functions used to evaluate PSO variations is presented in this study. Most of these functions were previously used in [4, 9, 31, 32]. These benchmark functions provide a trade-off between unimodal and multimodal functions.

Figure 2 shows the best solutions found by the PSO variations using the 14 benchmark functions. As mentioned previously the objective form using benchmark functions is to find the minimum solution and this depend on each benchmark [33], for example in the most of a benchmark the optimal value that close to Zero. On another hand, the optimal value for some benchmark is close to (- 450) like shifted benchmark functions. This is exactly shown in figure 2, all selection schemes try to be close to the optimal solution but TPSO got the first rank on the contrary EPSO got the worst solution, PPSO, GPSO and LPSO are respectively among them.

Tables II and III summarize the results of the PSO variations using the 14 benchmark functions in each convergence scenario, as shown in Table I. The results in Tables II and III are arranged from sen1 to sen5 to save the best value for each

parameter, which means in sen5 each of the selection schemes has the best values of parameters. Each PSO variation runs 30 replications, and the numbers in the table refer to the mean and standard deviations (within the parentheses below the mean value). The best solutions are highlighted in bold (i.e., the lowest is the best).

The results show that TPSO achieves the best results for all the benchmark functions. GPSO and PPSO achieve the eight best results for the Sphere, Schwefel problem 2.22, Step, Rosenbrock, Rotated hyper-ellipsoid, Rastrigin, Ackley, and Griewank benchmark functions. LPSO achieves the best results for most of the benchmark functions. By contrast, EPSO achieves poor results when compared with the other selection schemes, especially for the Rotated hyper-ellipsoid, Rastrigin, Shifted Sphere, and Shifted Rosenbrock benchmark functions.

V. CONCLUSION AND FUTURE WORK

This study proposed new variations of PSO based on different selection schemes. Each variation is a PSO incorporated with a selection scheme. The proposed PSO

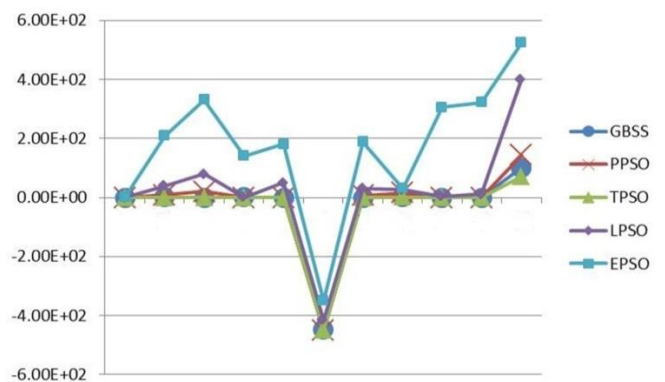


Fig. 2. Best solutions found by the PSO variations using the 14 benchmark functions

variations - GPSO, PPSO, TPSO, LPSO, and EPSO – employed the natural selection principle of the “survival of the fittest” to generate the new PSO. These variations focused on the better solutions to the solution space. The experiments were conducted with global benchmark functions that are widely used in the literature.

The experimental results show that incorporating the proposed selection scheme in the solution space by balancing exploration and exploitation prevents premature convergence and quick stagnation without efficient results.

This study also produced new four selection schemes: PPSO, TPSO, LPSO, and EPSO. The experimental results show that

these schemes perform better than GPSO. TPSO (the first position) achieves the best results, followed by PPSO and GPSO, whose results are close to each other. LPSO and EPSO are in the fourth and last positions, respectively.

This study is an initial exploration of selection schemes in the PSO algorithm. Future work should analyze these selection schemes in terms of takeover time [7]. The PSO performance in other benchmark and real-life problems should be investigated as well.

TABLE II. MEAN AND STANDARD DEVIATION OF THE BENCHMARK FUNCTIONS

benchmark function	Selection schemes	Sen1	Sen2	Sen3	Sen4	Sen5
Sphere	<i>GPSO</i>	6.97E-06 (9.99E-06)	4.27E-06 (2.21E-07)	1.47E-06 (1.92E-06)	3.11E-06 (8.77E-08)	2.80E-07 (9.85E-08)
	<i>PPSO</i>	9.50E-06 (4.17E-06)	8.23E-06 (2.45E-08)	6.58E-06 (4.25E-06)	4.87E-09 (1.983E-10)	2.50E-06 (4.17E-06)
	<i>TPSO</i>	0.00E+00 (0.00E+00)	6.93E-16 (2.21E-14)	0.00E+00 (0.00E+00)	0.00E+00 (0.00E+00)	0.00E+00 (0.00E+00)
	<i>LPSO</i>	3.38E+04 (9.48E+03)	4.25E-04 (3.50E-01)	4.80E-03 (7.30E-03)	9.87E-07 (8.08E-07)	5.57E-07 (1.23E-08)
	<i>EPSO</i>	3.25E-01 (1.09E-01)	3.24E-02 (2.11E-01)	1.56E-02 (1.88E-01)	2.43E-03 (1.02E-03)	1.47E-06 (1.92E-06)
Schwefel's problem 2.22	<i>GPSO</i>	0.0037 (0.0025)	0.0036 (0.0031)	7.99E-04 (6.65E-03)	7.88E-06 (2.43E-05)	8.43E-07 (1.12E-09)
	<i>PPSO</i>	0.0012 (0.0012)	5.85E-06 (0.0012)	4.76E-04 (3.77E-04)	2.44E-07 (8.75E-06)	9.23E-09 (4.65E-08)
	<i>TPSO</i>	4.33E-25 (0.00E+00)	4.57E-293 (0.00E+00)	2.57E-293 (0.00E+00)	0.00E+00 (0.00E+00)	0.00E+00 (0.00E+00)
	<i>LPSO</i>	0.1111 (0.0842)	7.35E-02 (6.84E-02)	9.76E-03 (5.90E-02)	1.02E-05 (8.83E-04)	5.23E-06 (3.87E-06)
	<i>EPSO</i>	2.08 E+02 (3.69 E+01)	2.07E+02 (3.81E+01)	1.004E+02 (4.259E+02)	0.92E+01 (3.87E+02)	0.69E+02 (0.54E+02)
Step	<i>GPSO</i>	7.28E-05 (6.50E-05)	9.50E-06 (1.20E-05)	6.21E-06 (3.57E-04)	8.94E-06 (2.55E-07)	1.75E-08 (2.54E-07)
	<i>PPSO</i>	1.51E-06 (2.06E-06)	1.51E-06 (2.06E-06)	2.07E-06 (2.72E-06)	6.63E-07 (1.94E-08)	8.78E-10 (6.59E-09)
	<i>TPSO</i>	0.00E+00 (0.00E+00)	0.00E+00 (0.00E+00)	0.00E+00 (0.00E+00)	0.00E+00 (0.00E+00)	0.00E+00 (0.00E+00)
	<i>LPSO</i>	8.20E-03 (1.23E-02)	8.20E-03 (1.99E-02)	5.69E-03 (1.59E+00)	1.22E-04 (6.82E-02)	4.98E-06 (2.73E-06)
	<i>EPSO</i>	3.38E+02 (9.47E+02)	3.19E+02 (1.00 E+02)	1.46E+01 (6.00E+01)	1.13E+01 (2.63E+01)	2.39E+00 (2.31E+00)
Rosenbrock	<i>GPSO</i>	0.03534 (1.9019)	1.22E-04 (1.38E-04)	2.55E-04 (7.69E-05)	6.64E-05 (4.33E-05)	1.65E-06 (2.37E-07)
	<i>PPSO</i>	1.31E-06 (2.26E-06)	8.31E-05 (2.26E-05)	6.58E-05 (4.25E-06)	2.35E-07 (4.67E-06)	3.08E-07 (6.29E-06)
	<i>TPSO</i>	0.00E+00 (0.00E+00)	0.00E+00 (0.00E+00)	0.00E+00 (0.00E+00)	0.00E+00 (0.00E+00)	0.00E+00 (0.00E+00)
	<i>LPSO</i>	0.6697 (2.2043)	3.41E-01 (1.59E+00)	3.85E-03 (5.21E-03)	4.32E-04 (8.94E-05)	2.82E-07 (7.65E-06)
	<i>EPSO</i>	1.46E+03 (6.70E+02)	3.28E+02 (6.00E+02)	1.52E+02 (2.11E+02)	2.17 E+01 (1.49E+02)	1.06 E+01 (2.00E+02)
Rotated hyper-ellipsoid	<i>GPSO</i>	5.55E-05 (9.06E-05)	9.20E-05 (3.24E-05)	5.55E-05 (9.06E-05)	5.69E-05 (2.44E-05)	5.59E-06 (2.39E-06)
	<i>PPSO</i>	7.61E-06 (1.23E-05)	7.61E-06 (1.23E-05)	3.651E-06 (7.32E-07)	6.84E-07 (7.56E-06)	5.69E-08 (3.49E-09)
	<i>TPSO</i>	0.00E+00 (0.00E+00)	0.00E+00 (0.00E+00)	0.00E+00 (0.00E+00)	0.00E+00 (0.00E+00)	0.00E+00 (0.00E+00)
	<i>LPSO</i>	1.16E-03 (0.0189)	2.39E-03 (1.09E+00)	1.04E-05 (3.40E-05)	1.56E-04 (5.21E-05)	3.85E-07 (5.21E-05)
	<i>EPSO</i>	1.80E+05 (6.56E+04)	1.49E+02 (3.29E+05)	3.24E+02 (2.11E+01)	2.33E+01 (2.83E-01)	6.94 E+00 (1.06E+01)
Schwefel's problem 2.26	<i>GPSO</i>	-448.576769 (2.095817)	-12564.817 (2.247382)	-12450.698 (2.948752)	-12557.657 (2.134256)	-12559.293 (2.267349)
	<i>PPSO</i>	-844.2568 (12.9384)	-2517.534 (656.294850)	-12558.592 (9.533928)	-12560.543 (2.434646)	-12563.685 (1.950643)
	<i>TPSO</i>	-930.816247 (257.157573)	-12567.43 (3.14213)	-12539.486 (0.000017)	-12539.493 (1.52E-03)	-12563.978 (1.08E-03)
	<i>LPSO</i>	-9754.924388 (399.855744)	-12561.42 (0.863088)	-12564.817 (2.247382)	-12553.343 (2.854832)	-12566.854 (1.098576)
	<i>EPSO</i>	-482.3852 (773.5379)	-9765.6465 (400.078376)	-9765.646 (400.078)	-9865.646 (241.532)	-11783.543 (223.495)
Rastrigin	<i>GPSO</i>	9.69E-03 (1.72E-03)	6.94E-03 (9.90E-04)	6.21E-05 (3.57E-04)	3.34E-06 (6.34E-05)	1.59E-06 (3.94E-06)
	<i>PPSO</i>	6.85E-06 (1.30E-06)	6.85E-07 (1.30E-06)	8.07E-07 (7.72E-07)	2.54E-07 (1.88E-07)	4.58E-08 (4.59E-07)
	<i>TPSO</i>	0.00E+00 (0.00E+00)	0.00E+00 (0.00E+00)	0.00E+00 (0.00E+00)	0.00E+00 (0.00E+00)	0.00E+00 (0.00E+00)
	<i>LPSO</i>	6.69E-02 (4.83 E-02)	1.39E-02 (1.39E-02)	5.69E-03 (1.59E-02)	3.01E-03 (8.22E-02)	4.45E-04 (2.15E-03)
	<i>EPSO</i>	1.88E+04 (2.65 E+04)	1.94 E+04 (4.19 E+04)	8.46E+03 (6.00E+04)	2.094E+03 (3.43E+03)	7.69E+01 (2.05E+01)

TABLE III. MEAN AND STANDARD DEVIATION OF THE BENCHMARK FUNCTIONS

benchmark function	Selection schemes	Sen1	Sen2	Sen3	Sen4	Sen5
Ackley	<i>GPSO</i>	5.65E-02 (4.26E-02)	2.46E-02 (5.34E-02)	1.01E-02 (6.54 E-03)	7.66E-03 (2.38E-04)	6.45E-04 (9.32E-04)
	<i>PPSO</i>	9.37E-03 (2.70E-02)	5.58E-04 (7.85E-04)	1.14E-04 (2.75E-04)	8.12E-04 (3.55E-05)	3.43E-06 (8.56E-07)
	<i>TPSO</i>	4.88E-03 (0.00E+00)	0.00E+00 (0.00E+00)	0.00E+00 (0.00E+00)	0.00E+00 (0.00E+00)	0.00E+00 (0.00E+00)
	<i>LPSO</i>	7.2318 (5.2197)	4.5235 (7.67E-01)	3.4935 (2.26E-02)	0.6697 (1.00E-02)	0.2849 (6.66E-3)
	<i>EPSO</i>	19.6590 (2.8761)	18.2253 (0.792)	18.2455 (7.29E-01)	17.4673 (3.32E-03)	13.9837 (1.27E-03)
Griewank	<i>GPSO</i>	6.35E-04 (1.38E-05)	2.38E-04 (5.89E-04)	1.65E-05 (7.08E-05)	3.23E-07 (8.63E-06)	4.54E-07 (7.35E-06)
	<i>PPSO</i>	2.38E-05 (5.89E-05)	5.80E-06 (7.44E-04)	5.80E-06 (7.44E-04)	6.85E-07 (8.30E-06)	7.26E-08 (5.68E-08)
	<i>TPSO</i>	0.00E+00 (0.00E+00)	0.00E+00 (0.00E+00)	0.00E+00 (0.00E+00)	0.00E+00 (0.00E+00)	0.00E+00 (0.00E+00)
	<i>LPSO</i>	0.0563 (0.0832)	6.06E-02 (6.09E-03)	7.03E-04 (1.86E-03)	1.25E-05 (7.43E-04)	1.02E-06 (2.45E-04)
	<i>EPSO</i>	1.34E+02 (1.76E+02)	4.96E+01 (6.65E+01)	3.87E+01 (2.97E+01)	3.02E+01 (2.97E+01)	2.11E+00 (1.02E+00)
Camel-Back	<i>GPSO</i>	-0.9275 (0.097)	-8.98E-01 (3.24E-01)	-9.18E-01 (4.84E-01)	-9.01E-01 (3.94E-01)	-8.67E-01 (3.26E-01)
	<i>PPSO</i>	-0.7673 (0.2075)	-7.98E-01 (5.012E-01)	-9.44E-01 (9.873E-01)	-9.95E-01 (9.86E-01)	-9.21E-01 (2.02E-01)
	<i>TPSO</i>	-0.7356 (0.00E+00)	-5.34E-01 (3.12E-01)	-9.79E-01 (7.02E-03)	-9.79E-01 (7.02E-03)	-9.98E-01 (9.87E-03)
	<i>LPSO</i>	0.7389 (2.2041)	1.53E+00 (3.63E+00)	1.01E+00 (1.03E+02)	1.01E+00 (1.03E+02)	-8.98E-01 (1.75E-02)
	<i>EPSO</i>	7.22E+02 (8.28E+02)	9.34E+02 (1.05E+03)	5.86E+02 (9.74E+01)	5.86E+02 (9.74E+01)	3.87E+01 (9.74E-01)
Shifted Sphere	<i>GPSO</i>	1.74E+03 (6.20E+02)	5801.689232 (1761.064344)	287.280860 (2085.974382)	-440.856 (2.642622)	-448.624 (1.364865)
	<i>PPSO</i>	1.85E+03 (6.10E+02)	-445.9264 (0.028288)	-447.6589 (0.757310)	-442.789 (1.45624)	-449.086 (2.456782)
	<i>TPSO</i>	2.64E+03 (1.03E+03)	-449.999836 (0.008747)	-447.999877 (0.000943)	-448.543 (1.43676)	-449.958 (0.076328)
	<i>LPSO</i>	3.27E+04 (8.94E+03)	5801.689232 (1761.064344)	3565.63289 (1076.69087)	753.538 (7.33E+04)	742.756 (2.39E+04)
	<i>EPSO</i>	1.94E+05 (4.90E+05)	5.80 E+04 (3.76 E+05)	2.57 E+04 (2.08 E+05)	9.56 E+03 (6.58E+04)	6.32 E+03 (6.65E+03)
Shifted Schwefel's problem 1.2	<i>GPSO</i>	5.41E+03 (1.26E+03)	946598.695 (309138.764)	-439.933552 (0.052206)	-440.863 (1.9564)	-449.661 (0.07654)
	<i>PPSO</i>	5.04E+03 (1.39E+03)	-48.748413 (366.028440)	-449.668252 (366.028440)	-441.698 (1.32E-01)	-449.827 (2.69E-02)
	<i>TPSO</i>	7.98E+03 (3.35E+03)	-447.007381 (2.969228)	-449.75496 (6.474446)	-448.365 (1.38E-02)	-449.947 (7.09E-02)
	<i>LPSO</i>	1.87E+05 (9.08E+04)	-449.933552 (0.052206)	-216.64485 (0.052206)	-421.302 (8.29E-02)	-439.546 (6.67E-02)
	<i>EPSO</i>	5.71E+03 (1.50E+03)	6598.695551 (9138.764095)	2643.859146 (7586.125628)	4085.025 (3878.464)	546.131 (558.315)
Shifted Rosenbrock	<i>GPSO</i>	2.12E+11 (2.24E+10)	515.19 (105.6652)	509.54320 (363.352)	501.3213 (206.653)	502.1478 (302.579)
	<i>PPSO</i>	2.14E+11 (2.60E+10)	497.01 (106.87)	498.744 (116614)	469.005 (2.86E+02)	465.744 (2.00E+02)
	<i>TPSO</i>	2.49E+11 (6.54E+10)	486.825 (120.596)	495.595 (123.986)	421.203 (332.845)	388.564 (203.432)
	<i>LPSO</i>	2.17E+12 (6.42E+11)	589.40 (309.94)	597.585 (109.454)	578.230 (283.865)	554.587 (229.545)
	<i>EPSO</i>	2.14E+11 (2.60E+10)	1506.80 (1.87 E+7)	1122.625 (1.56 E+6)	983.748 (1.72E+04)	876.432 (1.65E+04)
Shifted Rastrigin	<i>GPSO</i>	143.079 (26.8865)	-329.8838 (0.304559)	-329.235 (0.848076)	-302.738 (0.304559)	-319.454 (0.54786)
	<i>PPSO</i>	143.0297 (21.7146)	-329.951 (0.184239)	-429.768 (0.184239)	-320.765 (1.54677)	-323.564 (0.342)
	<i>TPSO</i>	143.6271 (17.9028)	-220.091 (13.0087)	-328.534 (2.87987)	-329.654 (4.51E-02)	-429.654 (3.69E-02)
	<i>LPSO</i>	8.79E+02 (172.0375)	-329.951 (0.184239)	-389.098 (1.760988)	-289.765 (0.3472)	-320.969 (0.7649)
	<i>EPSO</i>	923.7594 (47.333)	1026.852 (1876.38)	1751.450 (1555.89)	-201.543 (39.6543)	-281.203 (4.875)

REFERENCES

- [1] Z. Cui, and X. Gao, "Theory and applications of swarm intelligence," *Neural Computing and Applications*, vol. 21, no. 2, pp. 205-206, 2012.
- [2] M. Rabinovich, P. Kainga, D. Johnson *et al.*, "Particle swarm optimization on a gpu." pp. 1-6, 2012.
- [3] R. C. Eberhart, and J. Kennedy, "A new optimizer using particle swarm theory." pp. 39-43, 1995.
- [4] Y. Shi, and R. C. Eberhart, "Empirical study of particle swarm optimization", 1999.
- [5] X. Wu, Q. Xu, L. Xu *et al.*, "Genetic white matter fiber tractography with global optimization," *Journal of neuroscience methods*, vol. 184, no. 2, pp. 375-379, 2009.
- [6] R. Poli, J. Kennedy, and T. Blackwell, "Particle swarm optimization," *Swarm intelligence*, vol. 1, no. 1, pp. 33-57, 2007.
- [7] Y. Shi, and R. Eberhart, "A modified particle swarm optimizer." pp. 69-73, 1998.
- [8] N. Lynn, and P. N. Suganthan, "Distance Based Locally Informed Particle Swarm Optimizer with Dynamic Population Size." pp. 577-587, 2015.
- [9] R. C. Eberhart, and Y. Shi, "Comparing inertia weights and constriction factors in particle swarm optimization." pp. 84-88, 2000.
- [10] R. C. Eberhart, and Y. Shi, "Particle swarm optimization: developments, applications and resources." pp. 81-86, 2001.
- [11] S. Sen, P. Roy, and S. Sengupta, "Regular paper AI based Break-even Spot Pricing and Optimal Participation of Generators in Deregulated Power Market," *J. Electrical Systems*, vol. 8, no. 2, pp. 226-235, 2012.
- [12] M. A. Mohandes, "Modeling global solar radiation using Particle Swarm Optimization (PSO)," *Solar Energy*, vol. 86, no. 11, pp. 3137-3145, 2012.
- [13] E. Zahara, and Y.-T. Kao, "Hybrid Nelder-Mead simplex search and particle swarm optimization for constrained engineering design problems," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3880-3886, 2009.
- [14] C.-F. Juang, "A hybrid of genetic algorithm and particle swarm optimization for recurrent network design," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 34, no. 2, pp. 997-1006, 2004.
- [15] M. A. Al-Betar, I. A. Doush, A. T. Khader *et al.*, "Novel selection schemes for harmony search," *Applied Mathematics and Computation*, vol. 218, no. 10, pp. 6095-6117, 2012.
- [16] T. Back, "Evolutionary algorithms in theory and practice." Oxford Univ. Press, 1996.
- [17] J. H. Holland, "Adaptation in natural and artificial systems." *An introductory analysis with applications to biology, control, and artificial intelligence*: U Michigan Press, 1975.
- [18] J. J. Grefenstette, and J. E. Baker, "How genetic algorithms work: A critical look at implicit parallelism." pp. 20-27, 1989.
- [19] T. Blickle, and L. Thiele, "A Mathematical Analysis of Tournament Selection." pp. 9-16, 1995.
- [20] L. D. Whitley, "The GENITOR Algorithm and Selection Pressure: Why Rank-Based Allocation of Reproductive Trials is Best." pp. 116-123, 1989.
- [21] T. Blickle, and L. Thiele, "A comparison of selection schemes used in evolutionary algorithms," *Evolutionary Computation*, vol. 4, no. 4, pp. 361-394, 1996.
- [22] V. S. Gordon, and D. Whitley, "Serial and parallel genetic algorithms as function optimizers." pp. 177-183, 1993.
- [23] I. C. Trelea, "The particle swarm optimization algorithm: convergence analysis and parameter selection," *Information processing letters*, vol. 85, no. 6, pp. 317-325, 2003.
- [24] M. Richards, and D. Ventura, "Dynamic sociometry in particle swarm optimization." pp. 1557-1560, 2003.
- [25] U. KC, T. Deconinck, P. Varghese *et al.*, "Experimental and numerical studies of a direct current microdischarge plasma thruster." pp. 585-600, 2001.
- [26] Y. Shi, and R. C. Eberhart, "Fuzzy adaptive particle swarm optimization." pp. 101-106, 2001.
- [27] X. Jin, Y. Liang, D. Tian *et al.*, "Particle swarm optimization using dimension selection methods," *Applied Mathematics and Computation*, vol. 219, no. 10, pp. 5185-5197, 2013.
- [28] J. Kennedy, "The behavior of particles." pp. 579-589, 1998.
- [29] Z. Li-Ping, Y. Huan-Jun, and H. Shang-Xu, "Optimal choice of parameters for particle swarm optimization," *Journal of Zhejiang University Science A*, vol. 6, no. 6, pp. 528-534, 2005.
- [30] X. Hu, and R. Eberhart, "Solving constrained nonlinear optimization problems with particle swarm optimization." pp. 203-206, 2002.
- [31] J. J. Liang, A. K. Qin, P. N. Suganthan *et al.*, "Comprehensive learning particle swarm optimizer for global optimization of multimodal functions," *Evolutionary Computation, IEEE Transactions on*, vol. 10, no. 3, pp. 281-295, 2006.
- [32] P. Civicioglu, and E. Besdok, "A conceptual comparison of the Cuckoo-search, particle swarm optimization, differential evolution and artificial bee colony algorithms," *Artificial Intelligence Review*, vol. 39, no. 4, pp. 315-346, 2013.
- [33] J. Vesterstrom, and R. Thomsen, "A comparative study of differential evolution, particle swarm optimization, and evolutionary algorithms on numerical benchmark problems." pp. 1980-1987, 2004.

Knowledge Acquisition for Developing Knowledge-Base of Diabetic Expert System

IBRAHIM M.AHMED, MARCO ALFONSE, ABEER M.MAHMOUD, ABDEL-BADEEH M.SALEM

Computer Science Department

Faculty of Computer and Information Sciences

Ain Shams University, Cairo, Egypt

ibrahim1630@gmail.com, marco_alfonse@yahoo.com

abeer_f13@yahoo.com and abmsalem@yahoo.com

Abstract: - Diabetes is a serious health problem today. Most of the people are unaware that they are in risk of or may even have type-2 diabetes. Type-2 diabetes is becoming more common due to risk factors like older age, obesity, lack of exercise, family history of diabetes, heart diseases . Along with good lifestyle and healthy diet, reduces the risk of development of type 2 diabetes for treatment of elder people , proper care of diet, exercise and medication as well is more important.. The research in developing intelligence knowledge base systems in diabetic domain is important for both health industry and diabetes patients. Recently expert systems technology provides an efficient tools for diagnosing diabetes and hence providing a sufficient treatment. The main challenge in building such systems is the knowledge acquisition and developing of the knowledge base of these systems. Our research was motivated by the need of such an efficient tool. The main objective of this paper is gathering knowledge acquisition for developing the knowledge base of diabetic type-2 diet. Therefore, the paper presents the main phases of knowledge acquisition process on the development of fully automated healthy meal planner for diaptic-type-2.

Key-Words: *expert systems - semantic network, diabetic diet, type 2 diabetes, rule-base.*

1 Introduction

Diabetes is one of the major risky diseases for health care in our lives. If people were aware of the factors of diabetes and know how much risks they are of getting diabetes, diabetes may be prevented early [1]. Type 2 diabetes is a disease resulting from a relative, rather than an absolute, insulin deficiency with an underlying insulin resistance. Type 2 diabetes is associated with obesity, age, and physical inactivity [2, 3]. It is more common as compare to type-1 diabetes, usually 90 to 95%. It is diagnosed in both adults and young people. In this type pancreas does not produce enough insulin to control keeping blood sugar level within normal ranges. Actually it is serious type of diabetes where mostly people are not aware they are suffering from it. Three major causes of diabetes type 2 are lifelong bad diet, inactive or sedentary lifestyle, and overweight [4].

On the other hand the research in developing intelligence knowledge base systems in diabetic domain is important for both health industry and diabetes patients .Expert system is a computer program that provides expert advice as if a real person had been consulted where this advice can be decisions, recommendations or solutions. A few numbers of expert systems are utilized in diabetic health research where each of these systems attempts solving part or whole of a significant problem to reduce the essential need for human experts and facilitates the effort of new graduates [5].

The paper is organized as follows. Section 2 presents major risk factors Diabetic Diet and Diabetic Food Pyramid. Section 3 describes the related work, and in section 4 present the knowledge acquisition and representation process. Conclusion is given in the last section.

2 Diabetic Diet and Food groups

2.1 Diabetic Diet

Diabetic Diet for diabetics is simply a balanced healthy diet which is vital for diabetic treatment. The regulation of blood sugar in the non-diabetic is automatic, adjusting to whatever foods are eaten. But, for the diabetic, extra caution is needed to balance food intake with exercise, insulin injections and any other glucose altering activity. This helps diabetic patient to maintain the desirable weight and control their glucose level in their blood. It also helps to prevent diabetes patient from heart and blood vessel related diseases [6].

Research shows that regardless of the makeup of the diet, eating just enough calories to maintain an ideal weight is the most effective dietary strategy to prevent the onset of diabetic. Recommendations of diabetic diet differ for person to person, based on their nutritional needs, lifestyle, and the action and timing of medications. [7]

In Type 2 diabetic, the concern may be more oriented to weight loss in order to improve the body's ability to utilize the insulin it does produce. Thus, learning about the basic of food nutrition will be able to help in

adjusting diet to suite the particular condition. Recommended daily food portion contains carbohydrates, protein and fat. A Registered Dietitian assesses the nutritional needs of a person with diabetes and calculates the amounts of carbohydrate, fat, protein, and total calories needed per day. He will then convert this information into a recommended list of food for daily diet [7]. See table 1.

Table 1: Recommended daily food portion

Nutrition	daily calories
Carbohydrates	(50..55)%
Protein	(15..20)%
Fat	not more than 30%

2.2 Diabetic Food Pyramid

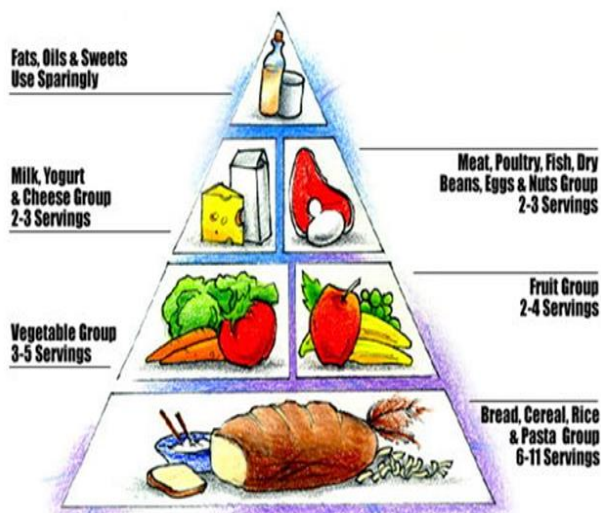


Figure 1: Food Pyramid

The Diabetes Food Guide Pyramid is a tool that shows how much you should eat each day from each food group for a healthy diet. The Diabetes Food Guide Pyramid is the best food guide for people with diabetes. The Diabetes Food Guide Pyramid places starchy vegetables such as peas, corn, potatoes, sweet potatoes, winter squash, and beans at the bottom of the pyramid, with grains. These foods are similar in carbohydrate content to grains. Cheese is in the Meat and others group instead of the Milk group because cheese has little carbohydrate content and is similar in protein and fat content to meat. [8].

Choosing foods from the Diabetes Food Guide Pyramid can help you get the nutrients you need while keeping your blood glucose under control [8].

Foods that are high in carbohydrates increase blood glucose levels and are in the Grains, Beans, and Starchy Vegetables group, the Fruits group, and the Milk group. Other foods that raise blood glucose are Sweets, found in the top of the Pyramid. Starchy foods, sweet foods,

fruits and milk are high in carbohydrate. Foods lows in carbohydrates are found in the Vegetables group, Meat and Others group and Fats. Diabetes patient should eat 6 to 11 servings Grains, 2 to 5 servings Group Vegetable, 2 to 4 servings Group Fruit, 2 to3 servings Group Milk, 2 to 3 servings group protein, Group sugars and oils should rarely be eaten [8].

2.3 Food groups

Food groups are exchange lists of foods that contain roughly the same mix of carbohydrates, protein, fat, and calories, serving sizes are defined so that each will have the same amount of carbohydrate, fat, and protein as any other. Foods can be "exchanged" with others in a category while still meeting the desired overall nutrition requirements. Food groups can be applied to almost any eating situation and make it easier to follow a prescribed diet. There are six food groups [9]:

1. Vegetables
2. Starches and Breads
3. Fruits
4. Milk
5. Fat
6. Meats and Meat Substitutes

The food groups are based on principles of good nutrition that apply to everyone. The reason for dividing food into six different groups is that foods vary in their carbohydrate, protein, fat, and calorie content. Each group contains foods that are alike; each food choice on a group contains about the same amount of carbohydrate, protein, fat, and calories as the other choices on that group [10].

3 Related work

M. Beulah et. al (2007) [11] introduced the ability to access diabetic expert system from any part of the world. They collect, organize, and distribute relevant knowledge and service information to the individuals. The project was designed and programmed via the dot net framework. The system allows the availability to detect and give early diagnosis of three types of diabetes namely type 1, 2, gestational diabetes for both adult and children.

W.Szajnar and G.Setlak(2011)[12] proposed a concept of building an intelligence system of support diabetes diagnostics, where they implemented start-of-art method based on artificial intelligence for constructing a tool to model and analyze knowledge acquired from various sources. The initial target of their system was to function as a medical expert diagnosing diabetes and replacing the doctor in the first phase of illness. Diagnostics the sequence of dealing with their system were as flow: (1) getting patient information and symptoms (2) competing basic medical examination in details (3) based on

previous information the system find out whether the patient has diabetes and decides whether it is type1 or type2. The systems used decision tree as a model for classification.

S. Kumar and B. Bhimrao (2012) [13] developed a natural therapy system for healing diabetic, they aim to help people's health and wellness, which don't cost the earth. Their main goal was to integrate all the natural treatment information of diabetes in one place using ESTA (Expert System Shell for Text Animation) as knowledge based system. ESTA has all facilities to write the rules that will make up a knowledge base. Further, ESTA has an inference engine which can use the rules in the knowledge base to determine which advice is to be

given to the user. Their system begins with Consultation asking the users to select the disease (Diabetes) for which they want different type of natural treatment solution then describes the diabetes diseases and their symptoms. After that describes the Natural Care (Herbal /Proper Nutrition) treatment solution of diabetes disease.

Bayu Adhi Tama, Rodiyatul F. S. And Hermansyah [14] proposed and boosted algorithm acquires information from historical data of patient's medical records of Mohammad Hoesin public hospital in Southern Sumatera. Rules are extracted from Decision tree to offer decision-making support through early detection of Type-2 diabetes mellitus for clinicians, table 2.

Table 2: Expert systems for diabetes

Authors	System purpose	ML technique	User interface
			Application
S.Kumar & B. Bhimrao 2012[13]	Integrate all the natural treatment information of diabetes in one place	rule based	Interactive
			Pc
W.Szajnar & Setlak 2011[12]	Model and analyze knowledge acquired from various sources	decision tree	Interactive
			Pc
Bayu.A.T et.al 2011[14]	Bayu.A.T et.al 2011[14] An Early Detection Method of Type-2 Diabetes Mellitus in Public Hospital	decision tree	Request /Response
			Pc
P. M. Beulah et.al 2007[11]	Detect and give early diagnosis of three types of diabetes for both adult and children	Rule based	Request /Response
			Pc

4. Knowledge acquisition and representation

4.1 Knowledge acquisition

Knowledge acquisition is a very important phase in developing expert systems [4]. Our knowledge has been gained by consultation of nutritionist. Actually, knowledge acquisition required time of three months

form major Ibtehal and Nasik nutritionist of diabetes in the military hospital in Khartoum, in addition to some related books and internet medical web sites. In addition we determine the amount of each item in the food groups in table 3.

Table 3: Standards of items

Fat & Milk		Sugar		Proteins	
Name	Amount	Name	Amount	Name	Amount
Oil	Spoon(20 gram)	Sugar	Spoon(20 gram)	Chicken	1/4 piece(250 gram)
Shortening	Spoon(20 gram)	Jam	Spoon(20 gram)	Egg	1 piece
Synths	Spoon(20 gram)	Cake	1 piece	Fish	125 gram
Milk	1 cup	Tahnia	Spoon(20 gram)	Meat	100 gram
Yogurt	100 gram	Sweet	1 piece	Tamiea	4 pieces(40 gram)
Cheese	50 gram	S_drinks	75 ml	Bean	100 gram
–	–	Basta	Small piece	Lentils	100 gram
–	–	–	–	Fual	100 gram
Fruits		Vegetables		Starch	
Name	Amount	name	Amount	Name	Amount
Banana	Small piece(100 gram)	Salad	Free	Custer	1 cup
Orange	Small piece(100 gram)	Molokhia	100 gram	Kissra	2 pieces(100 gram)
Mango	Small piece(100 gram)	Eggplant	100 gram	Gorasa	1/2 piece (100)
Dates	3 pieces(24 gram)	Dried Okra	100 gram	Bread	1 piece (120 gram)
Grapes	10 pieces (120 gram)	Potatoes	200 gram	Rice	1 cup
W_melon	2 slice(120)	Regala	200 gram	Pasta	1 cup
Apple	Small piece(100 gram)	Taglia	100 gram	Potato	Big piece
Guava	Small piece(100 gram)	Roub	200 gram	Noodles	1 cup

Kissra, Gorasa: kind of bread

4.2 Knowledge representation

Knowledge representation allows one to specify and emulate systems of a growing complexity. Knowledge representation schemes indeed have known an important evolution, from basic schemes supporting a rather heuristic approach, to advanced schemes involving

a deeper consideration of the various dependencies between knowledge elements [15]. The main Types of diabetes are Type1, Type2 and Gestational [16].figure 3 describes Knowledge representation of the diabetic serving.

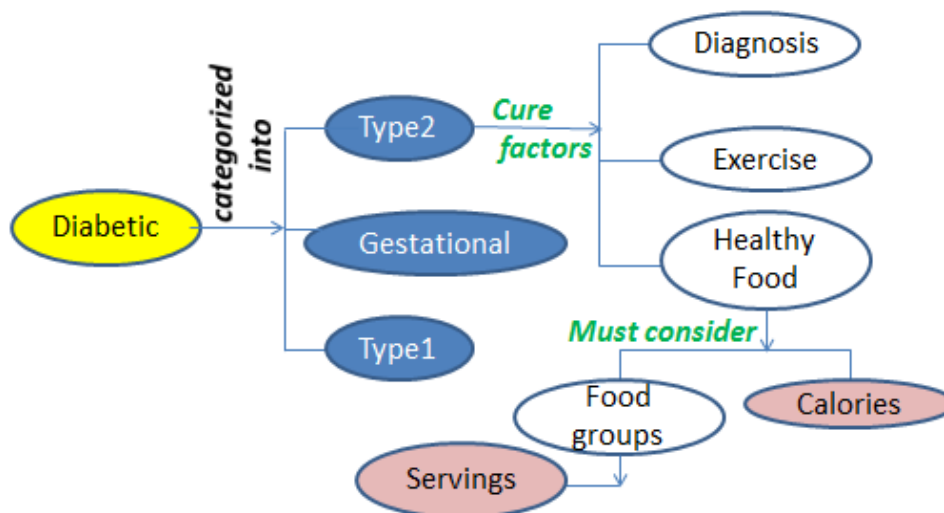


Figure 3: knowledge representation

4.3 Food groups servings

Some diseases increase the risk of diabetic disease and affect the number of serving in the food groups, the major diseases we get from our Knowledge acquisition are Anorexia, Surgery, Blood pressure, Typhoid, Bitter,

Liver problems, Heart disease and Gout. Other factors affect the serving are the patient activity, and weight see fig 4. The figure shows the relation of food groups and serving allowed for each group for the diabetic patient.

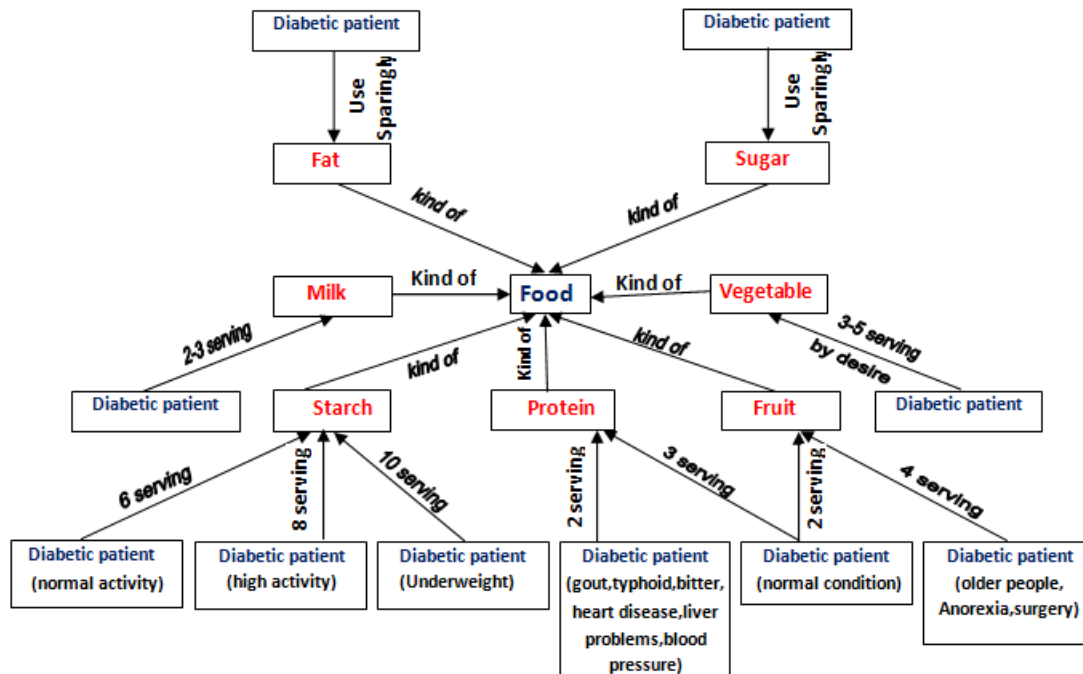


Figure 4: Semantic network for food groups servings

4.4 knowledge management

The following is the algorithm to specify the numbers of serving to each patient according to fig 4.

1. Determine whether the patient is slim or moderate or obese.
2. Determine whether the patient activity is high or moderate or little.
3. Determine whether the patient infected with (Anorexia, Surgery, Blood pressure, Typhoid, Bitter, Liver problems, Heart disease, Gout)
4. Calculate number of servings as follows:

Vegetable- servings =3

If (anorexia=1) or (surgery=1) or (age>65) then fruit-servings =4 else fruit- servings =2

If activity="normal" then crabs-servings=6

Else if activity="high" then crabs-servings=8

If the patient underweight then crabs-servings=10

If ((gout =1) or (Heart disease=1) or (Bitter=1) or (liver problems=1) or (Blood pressure=1) or (Typhoid=1)) then protein-servings=2 else protein-servings=3

If ((gout =1) or (Heart disease=1) or (Bitter=1) or (liver problems=1) or (Blood pressure=1) or (Typhoid=1)) then milk-servings=2 else milk-servings=3

5 Conclusions

Type-2 diabetes is the most common form of diabetes. This paper presents the first phase of developing an efficient expert system for diabetic Type-2 diet. The structure of the system contains three steps. First calculate total needs of calories, second determines the amount calories of the items and finally determines the proper diet.

Self-monitor for patient of type 2 diabetes is possible by getting proper amount of daily proper diet satisfy the amount of calories. The servings of meals calculate according to Body Mass Index (MBI) and the type of activity for the patient and the additional patient diseases. The food groups contain the same amount of carbohydrate, protein, fat, and calories Sudanese food

groups contains different meals so you don't have to eat the same foods all the time.

After collecting knowledge and perform the necessary analysis semantic network and food serving representation, Currently we are working on developing mobile-based expert system in Arabic language interface for diabetes diet that intended to be used in Sudan and Arab countries .

References:

- [1] Huiqing H. Yang and Sharnei Miller, "A PHP-CLIPS Based Intelligent System for Diabetic Self-Diagnosis", Department of Math & Computer Science, Virginia State University Petersburg, 2006.
- [2] Edward H. Shortliffe, Leslie E. Perreault, et al, Medical Informatics: Computer Applications in Health Care and Biomedicine, Springer-Verlag New York, Inc, 2001.
- [3] Federal Bureau of Prisons Management of Diabetes Clinical Practice Guidelines June 2012.
- [4] David Forbes, Pornpit Wongthongtham and Jaipal Singh. "Development of Patient-Practitioner Assistive Communications (PPAC) Ontology for Type 2 Diabetes Management", Curtin University, Perth, Australia, 2013.
- [5] Byoung-Ho Song, Kyoung-Woo Park and Tae Yeun Kim. "U-health Expert System with Statistical Neural Network", Advances on Information Sciences and Service Sciences. vol. 3, no.1, pp 54-61, 2011.
- [6](2013)[Online].Available:http://www.medmint.com/CONTENT/Diabetics/Diabetics_7.html Diet for diabetes patient.
- [7]Igbal.A and Nagwa. M,"health guide for diabetics", Sudan Federal ministry of health, 2010.
- [8] Mario A Garcia, Amit J.Gandhi, Tinu Singh, Leo Duarte, Rui Shen, Maruthi Dantu Steve Ponder, and Hilda Ramirez. "ESDIABETES (AN EXPERT SYSTEM IN DIABETES)", JCSC 16, pp 166-175. 2001.
- [9](2012)[Online].Available:[http://www.diabetes-diabetic-diet.com,-\"Diabetes Education and Prevention\" World Diabetes Day](http://www.diabetes-diabetic-diet.com,-\).
- [10](2013)[Online].Available:[http://www.glycemic.com/DiabeticExchange,\" The Diabetic Exchange List](http://www.glycemic.com/DiabeticExchange,\).
- [11] P. M. Beulah Devamalar, V. Thulasi Bai, and Srivatsa S. K. "An Architecture for a Fully Automated Real-Time Web-Centric Expert System", World Academy of Science, Engineering and Technology, 2007.
- [12] Wioletta SZAJNAR and Galina SETLAK. "A concept of building an intelligence system to support diabetes diagnostics", Studia Informatica, 2011.
- [13] Sanjeev Kumar and Babasaheb Bhimrao, "Development of knowledge Base Expert System for Natural treatment of Diabetes disease", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No. 3, 2012.
- [14] Bayu.A.T et.al," An Early Detection Method of Type-2 Diabetes Mellitus in Public Hospital ", 2 TELKOMNIKA, Vol.9, No.2, pp. 287-294, August 2011.
- [15] Stephan Grimm, Pascal Hitzler and Andreas Abecker," Knowledge Representation and Ontologies Logic, Ontologies and SemanticWeb Languages", University of Karlsruhe, Germany, pp 37-87, 2007.
- [16] Abdulla Al-Malaise Al-Ghamdi et al," An Expert System of Determining Diabetes Treatment Based on Cloud Computing Platforms", International Journal of Computer Science and Information Technologies, Vol. 2 (5), pp 1982-1987, 2011.

Solving Nurse Rostering Problem Using Artificial Bee Colony Algorithm

Asaju, La'aro Bolaji
Department of Computer
Science, University of Ilorin,
P.M.B. 1515 Ilorin, Nigeria.
lbasaju@unilorin.edu.ng

Mohammed A. Awadallah
Department of Computer
Science, Al-Aqsa University,
P. O. Box 4051, Gaza,
Palestine.
ma.awadallah@aldaqsa.edu.ps

Mohammed Azmi Al-Betar
Department of Information
Technology, Al-Huson
University College, Al-Balqa
Applied University, P. O. Box
50, Al-Huson, Irbid, Jordan.
mohbetar@bau.edu.jo

Ahamad Tajudin Khader
School of Computer Sciences,
Universiti Sains Malaysia,
Penang, Malaysia.
tajudin@cs.usm.my

Abstract—Artificial bee colony algorithm(ABC) is proposed as a new nature-inspired algorithm which has been successfully utilized to tackle numerous class of optimization problems belongs to the category of swarm intelligence optimization algorithms. The major focus of this paper is to show that ABC could be used to generate good solutions when adapted to tackle the nurse rostering problem (NRP). In the proposed ABC for the NRP, the solution methods is divided into two phases. The first uses a heuristic ordering strategy to generate feasible solutions while the second phase employs the usage of ABC algorithm in which its operators are utilized to enhance the feasible solutions to their optimality. The proposed algorithm is tested on a set of 69 problem instances of the dataset introduced by the First International Nurse Rostering Competition 2010 (INRC2010). The results produced by the proposed algorithm are very promising when compared with some existing techniques that worked on the same dataset. Further investigation is still necessary for further improvement of the proposed algorithm.

Keywords—Nurse Rostering; Artificial Bee Colony Algorithm; Swarm Intelligence Method; Nature-inspired algorithm

I. INTRODUCTION

The nurse rostering problem (NRP) is among the timetabling problem that is widely investigated by the researchers in the domain of operations research and artificial intelligence. The NRP as a NP-hard problem is described as an assignment of a set of qualified nurses to a different set of shifts over a predetermined scheduling period, subject to a set of hard and soft constraints. The hard constraints is the type that must be fulfilled for the roster to be *feasible* whereas the violations of soft constraints in the NRP are allowed but should be minimized as much as possible. It is noteworthy that the quality of the roster is determined by the satisfaction of the soft constraints in a feasible roster. The basic objective of NRP is to generate a feasible roster of high quality. However, studies in NRP domain have shown that it is almost impossible to find a roster that satisfies all constraints, since the NRP is classified as a combinatorial optimization problem [1] [2]. Due to the combinatorial and highly constrained nature of NRP, providing good quality roster is a very difficult and challenging task [3]. Naturally, investigations of numerous techniques for tackling NRP in the timetabling domain have increased over the past five decades. Some earliest techniques utilized for the NRP include integer programming [4], [5], goal programming [6], case-based reasoning [7], [8] and constraint Programming [9], [10]. In the recent time, some of the metaheuristic techniques that have also been employed for the NRP are local search-based approaches, which include tabu search [11], [12], simulated annealing [13], variable neighbourhood structures (VNS)

[14], [15]. Others are population-based approaches like ant colony optimization [16], genetic algorithm (GA) [17], [18], harmony search algorithm (HSA) [19], [20], particle swarm optimization [21]. Similarly, hyperheuristic and hybrid metaheuristic approaches have also been utilized [22]. The comprehensive review of the methodologies utilized in tackling NRP can be found in [23], [24], [25].

This paper tackles the NRP dataset proposed by the First International Nurse Rostering Competition (INRC2010), which is organized by the CODES research group at Katholieke Universiteit Leuven in Belgium, SINTEF Group in Norway and the University of Udine in Italy. The dataset of the INRC2010 is classified into three tracks: sprint, medium, and long datasets, which are varied in size and complexity. Each track is grouped into four categories in accordance with their publication time at the competition: early, late, hidden, and hint. Few techniques proposed to solve the INRC2010 dataset during and after the competition are review as follows.

Valoux et al. in [26] applied Integer Programming (IP) to tackle the NRP using INRC2010 dataset in which their solution method consists of two stages: the first stage consists of assigning different nurses to working days whereas the second stage involves scheduling of the nurses assigned to working days to certain shifts. The authors employed the use of three additional neighborhood structures in the first phase which are: (i) rescheduling one day in the roster for another time, (ii) rescheduling two days in the roster for another time, and (iii) reshuffling the shifts among nurses, for the medium and long track of the dataset.

The method ranked first in all three tracks of the dataset. The presentation of two methods to solve the INRC2010 dataset is presented in [27]. The authors in their work utilized the ejection chain-based method for the sprint track dataset while the branch and price method is employed for medium and long tracks of the INRC2010 dataset. The branch and price method achieved second rank for the medium and long tracks while the ejection method came fourth in sprint track of the dataset. The modeling of INCR2010 dataset as Constraint Optimization Problem (COP) is given in [28]. The author utilized the "COP solver" based on tabu search to further enhanced the results and the technique rated second, third, fourth in sprint, medium and long tracks of INRC2010, respectively.

Application of adaptive local search based on tabu search to tackle INRC2010 dataset is presented in [29] in which the solution method is also divided into two stages. The first phase involves the use of a random heuristic method to generates a feasible roster while the utilization of two neighborhood structures (i.e., move and swap) were employed to improve the solution at the second stage. It is worthy to mention that the method maintained the previous rosters in an elite pool. If the quality of the roster could not be improved within a given number of iterations with the aid of local search procedure, then one of the elite rosters is randomly chosen to restarts the second stage. The method achieved third and fourth position in the sprint and medium tracks respectively. The hybridization of a hyper-heuristic with a greedy shuffle move to solve INRC2010 dataset is presented by Bilgin et al. [30]. At initial stage, simulated annealing hyper-heuristic was employed to generate a feasible roster, where the satisfaction of soft constraints is achieved as much as possible. The greedy shuffle was used for further improvement of the roster. The hybrid hyper-heuristic method came third in long track, and fifth in sprint and medium tracks of the INRC2010 dataset. The introduction of a heuristic method for solving the INRC2010 dataset is considered in [31]. The heuristic method was employed in the construction of a feasible roster as well as trying to achieved the satisfaction of five pre-defined soft constraints. The authors utilized three local search procedures to further enhanced the roster. The method achieved the fifth position in long track. Adaptation of harmony search algorithm (HSA) was proposed for NRP using INRC2010 dataset in [19] where the results achieved on small instances of the dataset shows that method is very promising. The HSA was later modified in another development with inclusion of specific local search procedures in the pitch adjustment operator to minimized the violations of the soft constraints in [19]. The performance of modified HSA is further enhanced with the hybridization of greedy shuffle local search procedure which was utilized to enhance the new solution locally at each iteration [33]. Other HSA related works that have been employed to tackle NRP can be found in [20], [34], [35]. It is worth noting that research in the domain of NRP is still

active, since exact solution has as yet been found for the INRC2010 dataset which necessitated further investigations using other algorithmic techniques. The main purpose of this study is investigated whether the use of the Artificial Bee Colony Algorithm (ABC) could be utilized to improve the state-of-the-art results for the INRC2010 dataset.

II. NURSE ROSTERING PROBLEM

The Nurse Rostering Problem (NRP) could be solved by assigning a set of nurses with different skills and work contracts to a set of shift types over a given scheduling period. The solution (or roster) to NRP is subject to hard and soft constraints. The hard constraints in NRP must be fulfilled in the roster (i.e. H_1 and H_2 as shown in Table I). The fulfillment of soft constraints (i.e. S_1 - S_{10} see Table I) is desirable, and determines the quality of the roster. The basic objective is to find a roster that satisfies all hard constraints while minimizing soft constraints' violations. Table I shows the hard and soft constraints of INRC2010 datasets.

TABLE I
INRC2010 HARD AND SOFT CONSTRAINTS.

Hard Constraints	
H1	All demanded shifts must be assigned to a nurse.
H2	A nurse can only work one shift per day, i.e., no two shifts can be assigned to the same nurse on a day.
Soft Constraints	
S1	Maximum and minimum number of assignments for each nurse during the scheduling period.
S2	Maximum and minimum number of consecutive working days.
S3	Maximum and minimum number of consecutive free days.
S4	Assign complete weekends.
S5	Assign identical complete weekends.
S6	Two free days after a night shift.
S7	Requested day-on/off.
S8	Requested shift-on/off.
S9	Alternative skill.
S10	Unwanted patterns. (Where a pattern is a set of legal shifts defined in terms of work to be done during the shifts; Wren, 1996.)

Mathematically, the hard constraints for the INRC2010 dataset can be formulated as follows:

H_1 : All demanded shifts must be assigned to a nurse (see Eq. 1).

$$\sum_{i=1}^N x_i = d_{jk}. \quad (1)$$

H_2 : A nurse can only work one shift per day (see Eq. 2).

$$\sum_{i=1}^N x_i \leq 1. \quad (2)$$

Note that x_i is the allocation in the nurse roster (i.e. solution) (\mathbf{x}) assigned with a three elements (nurse u , day v , shift r). d_{jk} is the number of nurses required for day (j) at shift (k), where $v = j$, $r = k$, and N represents the maximum length of allocations for nurse roster (\mathbf{x}) as calculated in Eq. (3)

$$N = \sum_{i=0}^{W-1} \sum_{j=1}^7 \sum_{k=0}^{T-1} d_{((i \times 7) + j)k} \quad (3)$$

where W represents the total number of weeks in a scheduling period, T represents the total number of shifts.

The nurse solution (i.e. roster) is evaluated using an objective cost in Eq. (4), which sums up the penalty of soft constraint violations in a feasible roster.

$$\min f(\mathbf{x}) = \sum_{s=1}^{10} c_s \cdot g_s(\mathbf{x}) \quad (4)$$

It is noteworthy that s is the index of the soft constraint (S_1, \dots, S_{10}), c_s is the penalty weight for the violation of the soft constraint s , while $g_s(\mathbf{x})$ represent the total number of violations in \mathbf{x} for the soft constraint s , \mathbf{x} is a roster solution which represented as a vector shown in Fig 1

Fig. 1, Roster \mathbf{x} representation

x_1	x_2	x_3	x_{N-1}	x_N
Nurse 2	Nurse 5	Nurse 1	Nurse 11	Nurse 3
Day 5	Day 10	Day 9	Day 1	Day 19
Shift L	Shift E	Shift N	Shift D	Shift D

III. ARTIFICIAL BEE COLONY ALGORITHM

Artificial Bee Colony Algorithm (ABC) is a nature-inspired stochastic algorithm developed in 2005 by Karaboga in [36] for solving numerical problems. The ABC algorithm as a population-based search method is motivated by intelligent foraging behaviour of honey bee in their hives based on the model proposed in [37]. In ABC, the colony comprising three classes of bees: employed, onlooker, and scout bees. The colony is divided into two where the first half is occupies by the employed bees, while the second half consists of the onlookers bees. Each employed bee is associated with a solution (i.e., food source). In other words, the number of employed bees is equal to the number of food sources. The employed bee whose food source is abandoned by the onlooker automatically turns to a scout. The onlooker bees are those bees that hang around the hive to understudy the dance behavior of the employed bees in order to choose the desired solution. The scouts are ones that are randomly exploring the solution search space for new food sources. Normally, the number of food sources in ABC algorithm is equal to the number of solutions in the population.

Furthermore, the position of a food source signifies the possible solution for the optimization problem. The nectar amount of a solution (i.e., food source) represents the quality of the food source by that solution [36].

ABC algorithm has been applied and hybridized successfully for tackling real-world problems especially numerous formulations of the timetabling problems [38], [39]. A comprehensive review for ABC algorithm applications on several combinatorial optimization problems can be found in [40], [41].

A. Artificial Bee Colony for NRP

In this section, the concepts of Artificial Bee Colony Algorithm (ABC) as adapted for the NRP is discussed. The adapted ABC algorithm involves changing its continuous nature with integration of different neighbourhood structures in order to cope with the solution search space of NRP.

The nurse roster (i.e. solution) is represented as a vector of allocations $\mathbf{x} = (x_1, x_2, \dots, x_N)$ where each allocation contains three values (nurse, day, shift). For instance, let $\mathbf{x} = (1, 1, 1); (1, 2, 3), \dots, (5, 5, 4)$ be a feasible nurse roster. The roster is interpreted by ABC algorithm as follows: the allocation $x_1 = (1, 1, 1)$ means nurse n_1 is assigned to shift s_1 at day d_1 . The second allocation $x_2 = (1, 2, 3)$ means nurse n_2 assigned to shift s_2 at day d_2 , and so on. Note that representation of this roster is adopted in [20]. The description of six main procedural steps of ABC algorithm adapted for tackling NRP are given as follows:

1) *Initialization of ABC and INRC2010 parameters:* This step involves initialization of the three control parameters of adapted ABC that are needed for tackling the NRP: solution number (SN) which is the number of food sources in the population and similar to the population size in GA; maximum cycle number (MCN) which represents the maximum number of iterations; and limit that is responsible for the abandonment of solution, if there is no improvement for certain number of iterations and basically use in diversifying the search. Similarly, the NRP parameters that are drawn from the *INRC2010* dataset are also initialized. They are the set of nurses, the set of skill categories, the set of shift types, the scheduling period, the set of work contracts, matrix of weekly nurse demand, matrices of nurses preferences, and eventually the set of unwanted patterns. The job specification which includes: total number of shifts, minimum number of shifts, maximum number of consecutive working days, minimum number of consecutive working days, maximum number of consecutive free days, minimum number of consecutive free days, and maximum working weekend in four weeks.

2) *Initialization of the Food Source Memory:* The food source memory (FSM) is a memory allocation that consists

of sets of feasible food source (i.e. rosters) which is determined by SN as shown in Eq. 5 In this step, the feasible rosters are generated using the heuristic ordering approach and stored in ascending order in FSM according to the objective cost values that is $f(x_1), f(x_2), \dots, f(x_{SN})$. The function of heuristic ordering is to sorts the daily shifts in ascending order based on the level of difficulty. It noteworthy that the lowest weekly nurses demand is the most difficulty and thus, the required nurses of the ordered shifts will be scheduled starting with the most difficult and ending with less difficult one.

$$FSM = \begin{bmatrix} x_1(1) & x_1(2) & \dots & x_1(N) \\ x_2(1) & x_2(2) & \dots & x_2(N) \\ \vdots & \vdots & \ddots & \vdots \\ x_{SN}(1) & x_{SN}(2) & \dots & x_{SN}(N) \end{bmatrix} \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_{SN}) \end{bmatrix} \quad (5)$$

3) *Send the Employed Bee to the Food Sources:* in this step, the employed bee operator selects feasible nurse rosters sequentially from the FSM and exploits each roster using the neighbourhood structures to produce a new set of neighbouring solutions. The neighbourhood structures utilized by employed bee are:

- **Move Neighbourhood Structure (MNS):** The nurse of chosen allocation x_j is replaced with another nurse selected randomly to solve the violations of the soft constraint.
- **Swap Neighbourhood Structure (SNS):** The shift of selected allocation x_j is swapped with another shift on the same day for another selected allocation x_k .
- **Swap Unwanted Pattern (SUP):** This exchange a group of shifts among two nurses in which the chosen allocation x_j is replaced with another group of shifts on the same day for another chosen allocation x_k .
- **Token Ring Move (TRM).** The nurse of chosen allocation x_j is replaced by another nurse selected randomly, if the soft constraint S_7 is violated. Furthermore, the shift of a selected allocation x_j will be exchanged with another shift on which another nurse is working on the same day, for another selected allocation x_k to solve the violation of the soft constraint S_8 .

The fitness of each new roster is calculated. If it is better than that of candidate roster (i.e. food source), then it replaces the parent roster in FSM. This process is implemented for all solutions. The detailed of this process is given in Algorithm (1).

4) *Send the Onlooker Bees to the Food Sources:* Subsequent to the completion of employed bees exploitation process, the employed bees share the information of exploited food source (i.e. roster) with onlooker bees. The onlooker bees decide to follow certain employed bees and exploit their

corresponding food sources randomly using the set of neighbourhood structures discussed above based on proportional selection probability as shown in Eq. (6)

$$P_j = \frac{f(x_j)}{\sum_{k=1}^{SN} f(x_k)}$$

```

for  $i = 1 \dots SN$  do
   $i = RND()$  { $RND$  generates a random integer number
in range 1 - 4}
  if  $i = 1$  then
     $x^{i(new)} = MNS(x^i)$ 
  else
    if  $i = 2$  then
       $x^{i(new)} = SNS(x^i)$ 
    else
      if  $i = 3$  then
         $x^{i(new)} = SUP(x^i)$ 
      else
        if  $i = 4$  then
           $x^{i(new)} = TRM(x^i)$ 
        end if
      end if
    end if
  end if
  if  $x^{i(new)}$  is better than  $x^i$  then
     $x^i = x^{i(new)}$ 
  end if
next  $i$ 
end for

```

Algorithm 1: Employed Bee Phase

Note that the $\sum_{i=1}^{SN} p_i$ is unity

Thus, the roster with higher selection probability may be selected and adjusted to its neighbourhood using the same strategy as the employed bee. The fitness of the new roster is calculated and if it is better, then it replaces the current one.

5) *Send the Scout to Search for Possible New Food Sources:* Owing to continuous exploitation, some food sources may finally be exhausted in which they might be abandoned by its corresponding employed bee. Thus, the associated employed bee turns to a scout bee, and explores the solution search space randomly for a possible new food source to replace the abandoned one. Memorize the fitness of the best food source found so far in FSM.

6) *Stopping condition:* Repeat steps 3-5 until a stop condition is achieved, which is originally determined by *MCN*.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, the proposed adapted ABC for NRP is coded in Microsoft Visual C++ 6.0 on Windows 7 platform

on Intel 2.00 GHz Core 2.66 Quad processor with 2 GB of RAM. A dataset introduced by INRC2010 for nurse rostering is employed to evaluate the performance of the proposed ABC for NRP. The dataset is grouped into three tracks: sprint, medium, and long problem instances based on complexity and size. Each track of the competition is categorized into four types according to the publication time with reference to the competition: early, late, hidden, and hint. The sprint track comes with 33 problem instances that are classified into 10: early, 10: late, 10: hidden, 3: hint. They are the easiest, which comprises 10 nurses with one skill category and 3 to 4 different contract types, and the daily shifts are 4 for 28 days scheduling period.

In addition, the medium track contains 18 problem instances, which are grouped into 5: early, 5: late, 5: hidden, 3: hint. They are more complicated than the sprint track problem instances, which includes 30-31 nurses with 1 or 2 skills and 4 or 5 different contracts. The daily shifts are 4 or 5 shifts over 28 days scheduling period. Lastly, the long track includes 18 datasets, which are classified into 5: early, 5: late, 5: hidden, 3: hint. They are being referred to as hardest, which contains 49-50 nurses with 2 skills and 3 or 4 different contracts. The daily shifts are 5 shifts for 28 days scheduling period.

The parameters settings of the proposed adapted ABC are selected based on our preliminary experiments over NRP, where solution number (SN) is fixed at 10 while limit and MCN are set 100 and 10000 respectively. Table II shows the experimental results produced by the proposed technique for 69 problem instances of the INRC2010 datasets. The numbers in Table II refer to the penalty cost for the violations of the soft constraints (lowest is the best), which is computed based on the objective cost as shown in Eq. (4). Similarly, as shown in Table II, the best results obtained by the adapted ABC on 69 instances of INRC2010 are compared with those achieved by the six state-of-the-arts methods, which are listed as follows:

- T₁ Global Best Harmony Search Algorithm (Awadallah et al. [20])
- T₂ Adaptive tabu search with restart strategy (Lu and Hao [29])
- T₃ Integer programming with set of neighbourhood structures (Valouxis et al. [26])
- T₄ Variable Depth Search Algorithm and Branch and Price Algorithm (Burke and Curtois [27])
- T₅ Hyper-heuristic combined with a greedy shuffle approach (Bilgin et al. [30])
- T₆ Constraint Optimization Solver (Nonobe [28])

Basically, the proposed adapted ABC produced are very competitive results in comparison with those achieved by the six existing methods in all 69 problem instances of INRC2010 dataset. This is initial research of adapting ABC algorithm to the INRC2010 dataset. However, it is observed during run of experiment that the proposed methods suffers

stagnation in local optima as well as encountering premature convergence. These shortcomings shall be addressed in our next research.

V. CONCLUSION

In this paper, an adaption of Artificial Bee Colony Algorithm is presented for tackling the NRP. As the results have shown in Table II, the adapted algorithm is capable of solving nurse rostering problem. Although the results produced by the algorithm in this study are compared with

TABLE II
EXPERIMENTAL RESULTS OF ADAPTED ABC AND SOME
COMPARATIVE TECHNIQUES THAT WORKED ON INRC2010
DATASET

	Adapted ABC			T ₁	T ₂	T ₃	T ₄	T ₅	T ₆
	Best	Mean	Std.	Best	Best	Best	Best	Best	Best
Sprint_early01	62	63.9	1.9	58	56	56	56	57	56
Sprint_early02	64	65.3	1.1	60	58	58	58	59	58
Sprint_early03	58	61.6	3.9	53	51	51	51	51	51
Sprint_early04	66	67.9	1.8	62	59	59	59	60	59
Sprint_early05	63	63.6	0.5	59	58	58	58	58	58
Sprint_early06	58	59.7	2.0	56	54	54	54	54	54
Sprint_early07	61	62	0.8	58	56	56	56	56	56
Sprint_early08	58	62	3.7	57	56	56	56	56	56
Sprint_early09	58	60.4	2.2	57	55	55	55	55	55
Sprint_early10	57	59.7	2.2	53	52	52	52	52	52
Sprint_hidden01	44	46.2	1.3	41	32	33	-	-	-
Sprint_hidden02	38	42.8	5.0	35	32	32	-	-	-
Sprint_hidden03	71	74	3.1	70	62	62	-	-	-
Sprint_hidden04	76	77.6	1.8	79	66	67	-	-	-
Sprint_hidden05	65	68.4	3.0	62	59	59	-	-	-
Sprint_hidden06	161	182.6	15.8	202	130	134	-	-	-
Sprint_hidden07	178	193	18.1	196	153	153	-	-	-
Sprint_hidden08	245	252.3	7.5	266	204	209	-	-	-
Sprint_hidden09	371	379.6	9.3	273	338	338	-	-	-
Sprint_hidden10	327	344.7	19.1	346	306	306	-	-	-
Sprint_late01	49	51.9	3.2	45	37	37	37	40	37
Sprint_late02	52	53.5	1.6	49	42	42	42	44	42
Sprint_late03	56	58.5	3.0	55	48	48	48	50	48
Sprint_late04	89	100.5	6.6	104	73	75	75	81	76
Sprint_late05	53	53.8	1.0	51	44	44	44	45	45
Sprint_late06	47	48.1	1.5	43	42	42	42	42	42
Sprint_late07	52	57.8	5.5	60	42	42	42	46	43
Sprint_late08	17	23.8	8.3	17	17	17	17	17	17
Sprint_late09	17	26.2	5.4	17	17	17	17	17	17
Sprint_late10	56	57.5	1.6	54	43	43	43	46	44
Sprint_hint01	85	93.5	6.6	101	-	-	-	78	-
Sprint_hint02	57	59.9	2.6	59	-	-	-	47	-
Sprint_hint03	74	77.6	7.1	77	-	-	-	57	-
Medium_early01	260	266.9	5.3	270	240	240	244	242	241
Medium_early02	261	267	4.8	275	240	240	241	241	240
Medium_early03	259	267.4	6.2	265	236	236	238	238	236
Medium_early04	257	264.7	8.6	263	237	237	240	238	238
Medium_early05	329	333.6	6.5	334	303	303	308	304	304
Medium_hidden01	188	200.7	7.7	253	117	130	-	-	-
Medium_hidden02	284	298.1	10.9	361	220	221	-	-	-
Medium_hidden03	64	68	3.3	93	35	36	-	-	-
Medium_hidden04	100	105.4	4.1	135	79	81	-	-	-
Medium_hidden05	201	211.1	11.6	275	119	122	-	-	-
Medium_late01	206	223.4	11.6	254	164	158	187	163	176
Medium_late02	52	54.3	2.9	72	20	18	22	21	19
Medium_late03	70	72.6	2.1	75	30	29	46	32	30
Medium_late04	65	70.8	4.3	79	36	35	49	38	37
Medium_late05	178	192	11.9	238	117	107	161	122	125
Medium_hint01	69	77.2	7.9	89	-	-	-	40	-
Medium_hint02	141	151.3	10.8	194	-	-	-	91	-
Medium_hint03	187	225.6	42.5	242	-	-	-	144	-
Long_early01	242	247.2	3.8	256	197	197	198	197	197
Long_early02	277	284.1	5.2	299	222	219	223	220	219
Long_early03	269	277.4	5.0	286	240	240	242	240	240
Long_early04	337	346.6	8.3	356	303	303	305	303	303
Long_early05	327	332.1	5.3	337	284	284	286	284	284
Long_hidden01	445	457	11.2	747	346	363	-	-	-

Long_hidden02	130	132.4	2.1	225	89	90	-	-	-
Long_hidden03	59	62.4	3.5	121	38	38	-	-	-
Long_hidden04	47	50.9	2.8	134	22	22	-	-	-
Long_hidden05	76	81.5	4.0	146	45	41	-	-	-
Long_late01	288	296.9	6.5	601	237	235	286	241	235
Long_late02	293	311.5	30.7	596	229	229	290	245	229
Long_late03	306	311.1	5.8	585	222	220	290	233	220
Long_late04	303	313.7	9.6	621	227	221	280	246	221
Long_late05	141	151.5	9.6	393	83	83	110	87	83
Long_hint01	52	56.1	2.4	134	-	-	-	33	-
Long_hint02	39	44.7	8.8	102	-	-	-	17	-
Long_hint03	116	122.6	7.4	375	-	-	-	55	-

those produced by other state-of-the-arts techniques, which show the techniques is very competitive. In addition, the proposed algorithm produced comparable results on NRP, further improvement could still be made in the area of exploitation in order to enhance its performance while tackling the NRP. Therefore, our future work will focus on the enhancement of this technique in the following areas:

- To improve the adapted ABC by introducing more powerful and more structure local search mechanisms to handle specific soft constraints violations while tackling the NRP.
- To integrate adapted ABC algorithm with components of other metaheuristic algorithms.

REFERENCES

- [1] John J Bartholdi III, "A guaranteed-accuracy round-off algorithm for cyclic scheduling and set covering", *Operations Research*, vol. 29, no. 3, pp. 501–510, 1981.
- [2] Harvey H Millar and Mona Kiragu, "Cyclic and non-cyclic scheduling of 12 h shift nurses by network programming", *European journal of operational research*, vol. 104, no. 3, pp. 582–592, 1998.
- [3] Anthony Wren, "Scheduling, timetabling and rostering a special relationship? The Practice, and Theory of Automated Timetabling I: Selected Papers from 1st International Conference on the Practice and Theory of Automated Timetabling (PATAT I), Edinburgh, UK, 1996, pp 46-75.
- [4] Andrew J Mason and Mark C Smith, "A nested column generator for solving rostering problems with integer programming", in *International conference on optimisation: techniques and applications*, Curtin University of Technology Perth, Australia, 1998, pp. 827-834.
- [5] Broos Maenhout and Mario Vanhoucke, "Branching strategies in a branch-and-price approach for a multiple objective nurse scheduling problem", *Journal of Scheduling*, vol. 13, no. 1, pp. 77–93, 2010.
- [6] M Naceur Azaiez and SS Al Sharif, "A 0-1 goal programming model for nurse scheduling", *Computers & Operations Research*, vol. 32, no. 3, pp. 491–507, 2005.
- [7] Gareth Beddoe and Sanja Petrovic, "Enhancing case-based reasoning for personnel rostering with selected tabu search concepts", *Journal of the Operational Research Society*, vol. 58, no. 12, pp. 1586–1598, 2007.
- [8] Gareth Beddoe, Sanja Petrovic, and Jingpeng Li, "A hybrid metaheuristic case-based reasoning system for nurse rostering", *Journal of Scheduling*, vol. 12, no. 2, pp. 99–119, 2009.
- [9] Rong Qu and Fang He, "A hybrid constraint programming approach for nurse rostering problems", in *Applications and innovations in intelligent systems XVI*, Springer, London, 2009 pp. 211–224.
- [10] Haibing Li, Andrew Lim, and Brian Rodrigues, "A hybrid AI approach for nurse rostering problem", in *Proceedings of the ACM symposium on Applied computing (SAC 2003)*, Melbourne, Florida, 2003, pp. 730–735.
- [11] Kathryn A Dowsland, "Nurse scheduling with tabu search and strategic oscillation", *European journal of operational research*, vol. 106, no. 2, pp. 393–407, 1998.
- [12] Edmund Burke, Patrick De Causmaecker, and Greet Vanden Berghe, "A hybrid tabu search algorithm for the nurse rostering problem", in *Simulated evolution and learning*, Springer, Canberra, Australia, 1999, pp. 187–194.
- [13] RN Bailey, KM Garner, and MF Hobbs, "Using simulated annealing and genetic algorithms to solve staff scheduling problems", *Asia-Pacific Journal of Operational Research*, vol. 14, no. 2, pp. 27–43, 1997.
- [14] Edmund K Burke, Timothy Curtois, Gerhard Post, Rong Qu, and Bart Veltman, "A hybrid heuristic ordering and variable neighbourhood search for the nurse rostering problem", *European Journal of Operational Research*, vol. 188, no. 2, pp. 330–341, 2008.
- [15] Edmund Burke, Patrick De Causmaecker, Sanja Petrovic, and Greet Vanden Berghe, "Variable neighborhood search for nurse rostering problems", in *Metaheuristics: computer decision-making*, pp. 153–172. Springer, 2004.
- [16] Walter J Gutjahr and Marion S Rauner, "An aco algorithm for a dynamic regional nurse-scheduling problem in austria", *Computers & Operations Research*, vol. 34, no. 3, pp. 642–666, 2007.
- [17] Uwe Aickelin and Kathryn A Dowsland, "An indirect genetic algorithm for a nurse-scheduling problem", *Computers & Operations Research*, vol. 31, no. 5, pp. 761–778, 2004.
- [18] Margarida Moz and Margarida Vaz Pato, "A genetic algorithm approach to a nurse rostering problem", *Computers & Operations Research*, vol. 34, no. 3, pp. 667–691, 2007.
- [19] Mohammed A Awadallah, Ahmad Tajudin Khader, Mohammed Azmi Al-Betar, and Asaju La'aro Bolaji, "Nurse scheduling using modified harmony search algorithm", in *sixth International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2011)*, IEEE, Penang, Malaysia, 2011, pp. 58–63.
- [20] Mohammed A Awadallah, Ahmad Tajudin Khader, Mohammed Azmi Al-Betar, and Asaju La'aro Bolaji, "Global best harmony search with a new pitch adjustment designed for nurse rostering", *Journal of King Saud University-Computer and Information Sciences*, vol. 25, no. 2, pp.145-162, 2013.
- [21] Tai-Hsi Wu, Jinn-Yi Yeh, and Yueh-Min Lee, "A particle swarm optimization approach with refinement procedure for nurse rostering problem", *Computers & Operations Research*, vol. 52, pp.52-63, 2014.
- [22] Burak Bilgin, Patrick De Causmaecker, and Greet Vanden Berghe, "A hyperheuristic approach to belgian nurse rostering problems", in *Proceedings of the 4th Multidisciplinary International Conference on Scheduling: Theory and Applications (MISTA 2009)*, Dublin, 2009, pp. 693–695.
- [23] Brenda Cheang, Haibing Li, Andrew Lim, and Brian Rodrigues, "Nurse rostering problems - a bibliographic survey", *European Journal of Operational Research*, vol. 151, no. 3, pp. 447–460, 2003.
- [24] Edmund K Burke, Patrick De Causmaecker, Greet Vanden Berghe, and Hendrik Van Landeghem, "The state of the art of nurse rostering", *Journal of scheduling*, vol. 7, no. 6, pp. 441–499, 2004.
- [25] Jorne Van den Bergh, Jeroen Beliën, Philippe De Bruecker, Erik Demeulemeester, and Liesje De Boeck, "Personnel scheduling: A literature review", *European Journal of Operational Research*, vol. 226, no. 3, pp. 367–385, 2013.
- [26] Christos Valouxis, Christos Gogos, George Goulas, Panayiotis Alefragis, and Efthymios Housos, "A systematic two phase approach for the nurse rostering problem", *European Journal of Operational Research*, vol. 219, no. 2, pp. 425–433, 2012.
- [27] Edmund K Burke and Tim Curtois, "An ejection chain method and a branch and price algorithm applied to the instances of the first international nurse rostering competition, 2010", in *Proceedings of the 8th International Conference on the Practice and Theory of Automated Timetabling PATAT*. Citeseer, 2010, vol. 10, pp. 13 – 23.
- [28] Koji Nonobe, "Inrc2010: An approach using a general constraint optimization solver", Technical Report, INRC2010 (<http://kuleventortrijk.be/nrcpetition>).

- [29] Zhipeng Lu and Jin-Kao Hao, "Adaptive neighborhood search for nurse rostering", *European Journal of Operational Research*, vol. 218, no. 3, pp. 865-876, 2012.
- [30] Burak Bilgin, Patrick De Causmaecker, Benoît Rossie, and Greet Vanden Berghe, "Local search neighbourhoods for dealing with a novel nurse rostering model", *Annals of Operations Research*, vol. 194, no. 1, pp. 33-57, 2012.
- [31] Ademir Aparecido Constantino, Dario Landa-Silva, Everton Luiz de Melo, Candido Ferreira Xavier de Mendonça, Douglas Baroni Rizzato, and Wesley Romão, "A heuristic algorithm based on multi-assignment procedures for nurse scheduling", *Annals of Operations Research*, vol. 218, no. 1 pp. 165-183, 2014.
- [32] Mohammed A Awadallah, Ahamad Tajudin Khader, Mohammed Azmi Al-Betar, and Asaju La'aro Bolaji, "Nurse rostering using modified harmony search algorithm", in *Swarm, Evolutionary, and Memetic Computing*, Springer, India, 2011, pp. 27-37.
- [33] Mohammed A Awadallah, Ahamad Tajudin Khader, Mohammed Azmi Al-Betar, and Asaju La'aro Bolaji, "Harmony search with greedy shuffle for nurse rostering", *International Journal of Natural Computing Research (IJNCR)*, vol. 3, no. 2, pp. 22-42, 2012.
- [34] Mohammed A Awadallah, Ahamad Tajudin Khader, Mohammed Azmi Al-Betar, and Asaju La'aro Bolaji, "Hybrid harmony search for nurse rostering problems", in *Computational Intelligence in Scheduling (SCIS)*, 2013 IEEE Symposium, Singapore, 2013, pp. 60-67.
- [35] Mohammed A Awadallah, Ahamad Tajudin Khader, Mohammed Azmi Al-Betar, and Asaju La'aro Bolaji, "Harmony search with novel selection methods in memory consideration for nurse rostering problem", *Asia-Pacific Journal of Operational Research*, vol. 31, no. 3, pp. 1-39, 2013.
- [36] D. Karaboga, "An idea based on honey bee swarm for numerical optimization", *Techn. Rep. TR06*, Erciyes Univ. Press, Erciyes, 2005.
- [37] D. Teodorović and M. Dell'Orco, "Bee colony optimization - a cooperative learning approach to complex transportation problems", in *Advanced OR and AI Methods in Transportation. Proceedings of the 10th Meeting of the EURO Working Group on Transportation*, Poznan, Poland. Citeseer, 2005, pp. 51-60.
- [38] Asaju La'aro Bolaji, A.T. Khader, M.A. Al-betar, and M. Awadallah, "The effect of neighborhood structures on examination timetabling with artificial bee colony", in *Practice and Theory of Automated Timetabling IX*, Son, Norway, 2012, pp. 131-144.
- [39] Asaju La'aro Bolaji, Ahamad Tajudin Khader, Mohammed Azmi Al-Betar, and Mohammed A Awadallah, "University course timetabling using hybridized artificial bee colony with hill climbing optimizer", *Journal of Computational Science*, vol. 5, no. 5, pp. 809-818, 2014.
- [40] D. Karaboga, B. Gorkemli, C. Ozturk, and N. Karaboga, "A comprehensive survey: artificial bee colony (abc) algorithm and applications", *Artificial Intelligence Review*, vol. 42, no. 1, pp. 1-37, 2012.
- [41] A.L. Bolaji, A.T. Khader, M.A. Al-Betar, and M.A. Awadallah, "Artificial bee colony, its variants and applications: a survey", *Journal of Theoretical & Applied Information Technology (JATIT)*, vol. 47, no. 2, pp. 434-459, 2013..

Towards Developing an Intelligent HAJJ Guide system

PILGRIM TRACKING AND IDENTIFICATION USING MOBILE PHONES

Malak Osman Abbdelazeez

Sudan University of Science and Information Technology-Sudan
Khartoum, Sudan
angel_osman@yahoo.com

Adnan Shaout

The University of Michigan – Dearborn
The Electrical and Computer Engineering Department
Dearborn, MI, USA
shaout@umich.edu

Abstract— This paper presents the development of a system designed to track and identify pilgrims in the Holy areas in Makah-Saudi Arabia during the Hajj season (Pilgrimage). The target area is covered by a complicated network through several service providers. Mobile phone sends UID, latitude, longitude and time stamp frequently or as requested. On Google map there is a server that maps the longitude information and latitude. In case the internet connection is lost then the mobile phone saves the location information in its memory until the connection is restored. Once connection is restored, it will then send the saved location information to the server and clears this information from memory. The developed system works collaborating with a GPS identification system.

Keywords— *Tracking System, Identification system, Hajj, GPS*

I. INTRODUCTION

It is well known that Hajj (Pilgrimage) is the most annual crowded Muslim gathering event on earth. The annual event has distinctive characteristics such as the religious practices (rituals) performed, the people who attend (pilgrims) and the place they meet in. These characteristics cause difficult challenges to the authorities in order to control the crowd and identify the personalities. What increases the challenge is the unity of movements as they all move from one place to another at the same time practicing the same rituals. Although a Muslim should perform Hajj once in a lifetime many prefer to perform it more. It is performed on the 8th-13th days of the 12th Hijri month in fixed boundaries around Makah city in Saudi Arabia. Although the Hajj authorities try their best to limit the flood of the crowd to the area by assigning quotas for pilgrims to each country, the number of pilgrims is still exceeds 2.5million annually and the number keeps growing.

The Saudi Ministry of Hajj released a statistic report stating that the number of pilgrims in 2014 was 2,085,238, where 73% of pilgrims are non-Saudi pilgrims [1]. A number of around 4 million pilgrims who come to the Holy areas every year other than Hajj times may benefit from the developed system as well. It is expected that the number of visitors will reach 10 million every year in the near future. Despite being a great spiritual experience for all pilgrims, at the same time it poses great challenges of all sorts for the authorities

responsible for facilitating the Hajj. Koshak et al. [2] mentioned that apart from the Hajj period, Makah areas become very crowded during the last ten days of Ramadan. During the month of Ramadan, it was reported that there are more than 2,500 cases of missing people in the area of Masjid al-Haram, the grand mosque in Makah [3]. Adopting such a worrying figure even before the Hajj period begins would be very dangerous and if no further improvements are made, the safety and security of the pilgrims would be jeopardized. In spite of all that is done to facilitate the Hajj, some common difficulties are facing the pilgrims and the authorities which are listed as following:

- ✓ Identification of pilgrims (lost, dead, or injured)
- ✓ Medical Emergencies
- ✓ Guiding lost pilgrims to their camps.
- ✓ Crowd control

The aim of this paper is to study the Hajj pilgrimage crowd tracking and identification problem as well as propose a better Hajj pilgrimage tracking and identifying system that is reliable and affordable. The paper is organized as follow: section 2 presents related work; section 3 presents Hajj locator architecture; section 4 presents the hardware specifications for the Hajj locator system; section5 covers the software specifications of the proposed system; section 6 presents the

system design; and section 7 presents the proof of concept model for the proposed system.

II. RELATED WORK

There has been quite a number of tracking and monitoring systems developed for crowd management. Each system uses its own means and facilities to increasing its effectiveness. One of the most widely recognized system is the tracking via RFID chips. Nowadays, a lot of embedded RFID chips are placed in our belongings and because of its relatively small size; it has been used quite extensively for many applications. In order to have a system that suits events such as Hajj, Yamin et al. [4] proposed to track people using the RFID chip and wireless technologies which uses a database to save data and the entities for each person. Installing sensor networks for sensing and reading the chips for irregular events does have some serious economic considerations. Another approach is having an object recognition system where a picture, which usually is a land object or structure, is taken using a built-in camera common in any mobile phone to identify their location with respect to the picture taken [5]. GPS is used to read the actual location if available. If the data cannot be acquired then it uses an approximate evaluation of the cell information of the phone-network provider. As good as the system might get, the system relies solely on Internet connectivity. People need to register to have their own Internet connection available in their mobile phones. For those who do not have Internet connection, it is burdensome to go through the requirements only to be used for a short period time during Hajj. Another approach is by implementing a low cost object tracking system using GPS and GPRS [6]. The system allows a user to view the present and the past positions recorded of a target object on Google Maps through the Internet. It reads the current position of the object using GPS, the data then is sent via GPRS service from the GSM network towards a web server. Some might argue that using SMS is an expensive means of communication. Although it is cheaper to use wireless network technologies when usage is heavy, it is expensive when consider the duration of time it is used. Another approach is a prototype using passive RFID technology passed through several implementations and discussions with Hajj officials [7]. The developed prototype was tested on 1000 pilgrims from the country of Ivory Coast in collaboration with officials from the Hajj Ministry. The results of the experiment have convinced the Hajj authorities to utilize this technology for all pilgrims in the near future. However, authorities indicated the need for tracking pilgrims in addition to the identification process. Therefore, an active RFID system is developed for tracking pilgrims to work on coordination with the passive RFID system for identification. However, the system faced several difficulties and proved to be impractical, particularly with the crowd. Thus the idea of using wireless sensor network for tracking pilgrims was introduced [8, 9]. The tracking and monitoring system consists of portable wireless sensor units carried by the pilgrims and a fixed Wireless Network (WSN) infrastructure capable of gathering,

processing and routing location and time stamp data of sensor units carried by the pilgrims. All the nodes in the fixed WSN are made equivalent to keep the deployment, configuration and reconfiguration process simple. Table 1 presents a comparison of the work that has been cited in literature regarding Hajj mobile applications.

Table 1: Presents a comparison of the work that has been cited in literature

Author	Solution	Economic Considerations	Internet Connectivity	Register
Yamin et al [4]	to track people using the RFID chip and wireless technologies	Have some serious economic considerations.		People need to register.
Luley, P et al [5]	having an object recognition system where identify their location according to the picture taken		The system relies solely on Internet connectivity.	People need to register to have their own Internet connection.
Hasan, K.S [6]	The system allows a user to view the present and the past positions recorded of a target object on Google Maps through the Internet.	Low cost object tracking system using GPS and GPRS.	Although it is cheaper to use wireless network technologies when usage is heavy, it is expensive to use if we consider the duration of time it will be used in, with the money we pay for it	
M. Mohandes [7]	Using passive RFID technology passed through several implementations.	The system faced several difficulties and proved to be impractical, particularly with the crowd.	Consists of portable wireless sensor units carried by the pilgrims and a fixed Wireless Network Infrastructure capable of gathering, processing, and routing data on locations and time stamps of sensor units carried by the pilgrims.	

regarding Hajj mobile applications.

During Hajj the pilgrims will only be there for around a month and getting Internet services from ISPs might be troublesome and therefore might result in inability to use the local service. Also it was noticed that an expensive infrastructure would need to be built and the cost of each portable sensor unit is not significantly cheaper than a mobile phone equipped with a GPS unit. Particularly that the majority of pilgrims have their own mobile phones and a large percentage of their phones are already equipped with GPS. This leads us to develop a pilgrim tracking system using mobile phones. The System which is proposed promotes accessibility by choosing a common platform that is widely used by people, which is the mobile phone. The proposed system also provides connections availability towards the user, where we use two types of connections in updating the data to the server. We believe that the proposed system once implemented will provide a better way of tracking pilgrims.

III. HAJJLOCATOR ARCHITECTURE OVERVIEW

The framework of the proposed tracking system is designed in two parts; the mobile device of the connection server and the tracking system using the database server as shown in Figure 1. The GPS-enabled mobile phone is connected with the user and the coordinates are updated in the server and stored in the database server. The coordinates then are sent with the Subscriber Identity Module (SIM) card number as its identification together with other useful data.

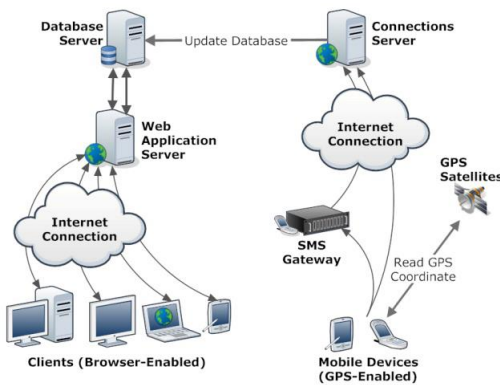


Figure 1: The mobile device of the connection server and the tracking system using the database server

The server which only provides reliable indoor and outdoor user location is divided into three parts; server side, processing side and connection side. Dissolution of servers is needed to handle the huge clients' updates. The database should be divided into different servers to ease the process of updates and avoids bottleneck.

As for of connectivity, different services offered by any GSM mobile phone can be utilized. Any wireless network infrastructures available can be used, together with SMS as the

means of data communication among clients and servers. The main precedence will be given to updates using any available Internet connection such as Wi-Fi, GPRS and 3G were it will then make use of the connection to update the server with pilgrim's GPS coordinates. In addition to that, SMS is also used as the other alternative connection to update the server. If Internet connections are not available, the device will then automatically use SMS as another option. This works as a solution for the availability issue especially in a situation such as alerting for a missing pilgrim .Security is also considered for this proposed tracking system. In case the administrative or authorized personnel wants to trace pilgrims, they need to log into the web server and get the position of the pilgrim using one of two choices; in a Google Maps view, and in a tabular view. Regarding security concerns, it is considered as control privacy, thus we will authenticate any user who wants to access the data.

A. Hardware Specifications

A HTC Touch Diamond2 smart phone has been selected in implementing the prototype for the Hajj Locator system. The Hajj locator is modeled as a stack hardware system as shown in Figure 2. The phone uses Windows Mobile 6.1 Professional and has an internal GPS antenna. In terms of network, it supports HSDPA/WCDMA network of up to 2 Mbps up-link and 7.2 Mbps down-link speeds, a Quad-band GSM/GPRS/EDGE and a Wi-Fi IEEE 802.11. More information about the mobile phone can be accessed through the web site at [10]. The server is running under normal PC with AMD Phenom 9600B Quad-Core Processor 2.31 GHZ, 1.75 GB of RAM.

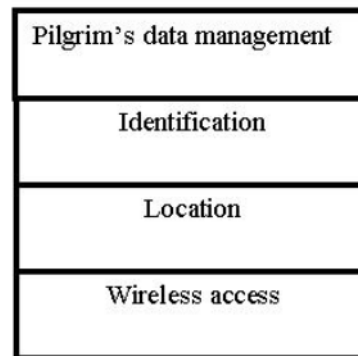


Figure 2: The hardware system stack

B. Software Specifications

For server update, a client application has been designed in parallel with the server process. We used java to implement the application. It reads the latitude and longitude of the location and process it based on the specifications defined by the user. The main specifications are through the distance-based and time-based parameters. The choice of distance-based and time-based parameter is designed to offer flexibility to the user in updating to the server. Figure 3 shows the flow diagram of the

distance-based and time-based latitude and longitude calculation updates in the server. For the distance calculation of coordinates, the Cosine-Haversine formula technique has been used [11]. It results in a great-circle distance between two points on a sphere given the latitudes and longitudes. However, other parameters can also be used manually according to the user request such as the “Mark My Location” and the “panic alert”. Mark My Location is a special button designed for users who intended to update the server about their current location. It uses only the chosen connectivity method to communicate with the server. This data of latitude and longitude will help the system to store reference points. The panic button is designed to alert the system in emergency situations. It uses all of the available resources, i.e. Wi-Fi, GPRS and SMS, to update current location of the user.

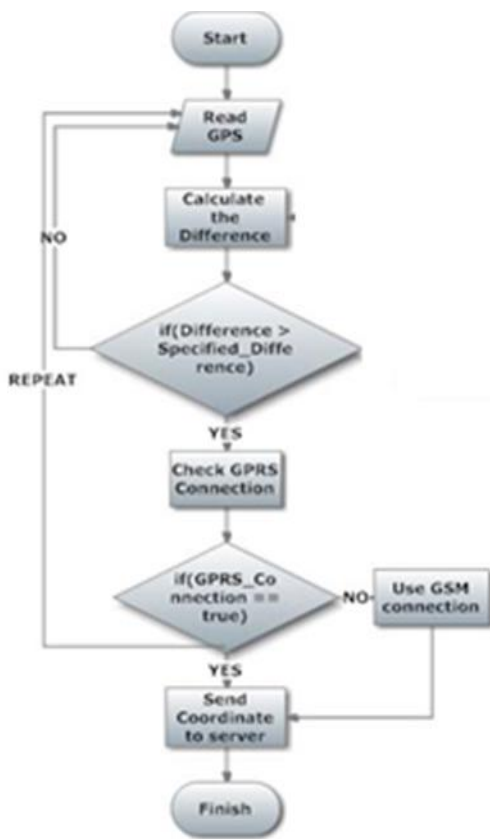


Figure 3: The flow diagram of the distance-based and time-based latitude and longitude calculation updates in the server.

In order to compensate the real-time update and at the same time being cost-effective, we have decided to use the dynamic update triggering to the server. In other words, the user will be in an area of circle with a defined radius and the system will only send an update to the server, should the user moves more than the defined distance. For this reason, to gain popularity and widespread usage, our proposed System is developed in a balanced approach and provides the ability to facilitate real

time update in a cost-effective way. On the server side, the development was done using, Java, AJAX and DHTML. To provide reliable data management, MYSQL is used in database server. To read coordinates from database, we use a PHP file, and parse it into an XML format. These XMLs are then will be processed by the application processing server. To display the tracking and monitoring of the user, a web based application has been developed. Through the web application, administrative or authorized personnel will be able to view the live position of the tracked user, together with the past positions and the route they have chosen. A web application is responsible for accepting data that has been sent by the mobile device via GPRS or GSM, using GET method of the HTTP protocol. This data consists of SIM number of the device, latitude, longitude, time, date, update mode, and distance between two consecutive coordinates based on their updating mode. SIM number is used to authenticate the device. For the real-time aspect, we use the technique of checking the database in periodic basis. Once real-time mode is activated, the current time will be stamped and then the database will be checked, as shown in Figure 4. If new data was found, the marker will be added to the map. Checking the database in a specific interval will automatically animate the marker on the map. We use the publicly accessible Google Maps API for some part of the code.

```

function realTimeUpdate()
{
    GDownloadUrl(sqlXmlUrl, function(data)
    {
        var xml = GXml.parse(data);
        markers =
        xml.documentElement.getElementsByTagName(
        "marker");
        var databaselatestdate =
        markers[0].getAttribute("date");

        var latestdate =
        changeToDate(databaselatestdate);
        if( latestdate >= currentTime)
        {
            currentTime= new Date();
            addMarkerToMap(0);
        }
    });
}
    
```

Figure 4: The code for the function for real-time update calculations.

IV. SYSTEM DESIGN

The developed system uses web service as the back end and mobile application is used to obtain the location information and sends it to the web service. The web service saves the received data in a database server using a secure channel and then the website connects to the web service to retrieve a specific user location and show his/her location on a Google map. The mobile application is developed using Java. The application has the following tasks:

- Obtain the current user’s location, and
- Based on a predefined time parameter send the location data (longitude, latitude, and time stamp)to the web service

The mobile application continues sending the location data periodically until the administrator stops the process or closes the application. The application uses assisted GPS. The location data is sent using Internet provided by either wireless network that is available in the Holy sites during Hajj season or using GPRS over HTTP protocol and using SOAP for data exchange with the server. If location information is not sent for any reason, it will be saved in the memory of the mobile until connection is restored. At that stage all gathered location information with the time stamps are sent and the information is cleared form the mobile memory. If the location information data size exceeds a pre-specified limit, the old location data is cleared to free the memory for new location information. Web service is used as an interface for both the mobile application and the website. It provides a web method to save the current user location using the mobile application and also retrieve the data either for a specific pilgrim or a group of pilgrims. The web service is implemented using PHP and it connects to a back end database exchange data between the mobile application and website. It connects to the web service to retrieve specific pilgrim’s locations and then it uses Google static APIs to show these locations on a Google map. The database is used to store and retrieve the user’s data and it is accessed only from the web service secured by authenticated administrator. The database is implemented using MYSQL and it is hosted on a separate server. Figure 5 shows a snap shot of the GU for the system on a mobile phone.

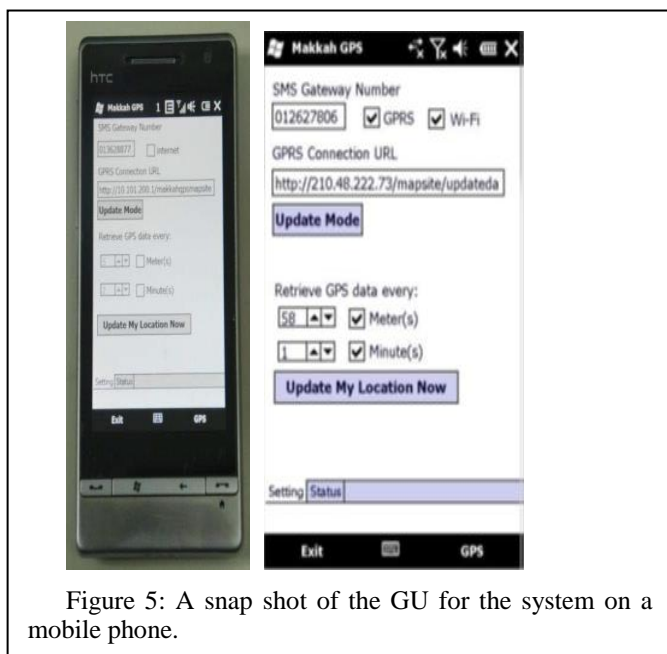


Figure 5: A snap shot of the GU for the system on a mobile phone.

The following section describes the software tools and platforms that were used to the development the system:

- Programming languages used:
 - Java for development for the mobile application
 - PHP for web Service and website development
 - MYSQL
 - ✓ Tools
 - Symbian Series 60 3rd edition SDK“S60-SDK-200634-3.1-Cpp-f.1090b”
 - Nokia PC Suite
 - ✓ Website and Web service
 - PHP XML
 - Database
 - MYSQL

A concern for pilgrim identification system using RFID is the tag type, format and quality. One can use a wristband tag or a plastic card tag that is hanged on the neck of the pilgrim as a business card. In the first case there was a concern that during the cleaning before prayers (Wudo) were the pilgrim has to clean his/her hands until the elbows and so he/she may take off the wristband and thus may lose it. In the latter case some pilgrims may not feel comfortable putting the card on their neck and thus may lose it. With the proposed system of using mobile phone for tracking, the RFID chip can be placed inside the mobile phone. Users already are extremely careful not to lose their mobile phones. Moreover, we have started working on using mobile phones with NFC (Near Field Communication) capability so that the same mobile phone can be used for tracking as well as identifying and thus all what a pilgrim needs during Hajj journey is his/her mobile phone. Figure 6 shows a high level design for the proposed system.

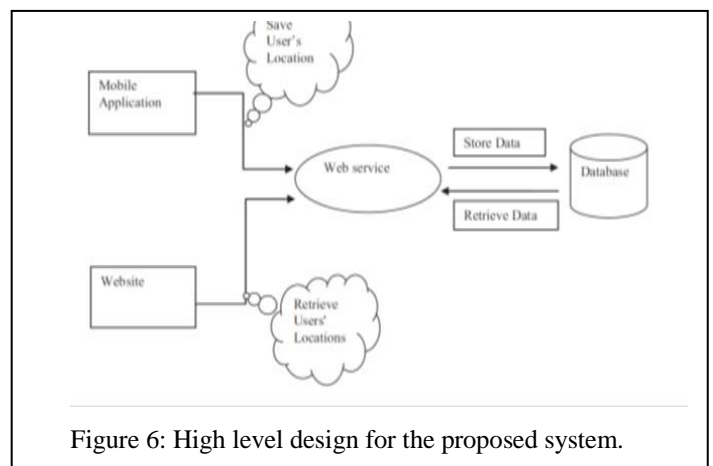


Figure 6: High level design for the proposed system.

V. PROOF OF CONCEPT

The developed system was tested in Makah, the Holy sites, during the 2014 Hajj season. A set of pilgrims having mobile

phones equipped with GPS units volunteered to test the system. The developed system was downloaded into the mobile phones and preset to get the location information and send it every minute. It was very easy to change the time interval between sending the successive location information. Each volunteering pilgrim was given the option to start the operation time so that he/she can collect and send the information in the Holy area and during the pilgrimage days only. During the four pilgrimage days the location information was collected and sent from all participating phones automatically without the disturbance or distraction of the pilgrims. The system run on the back ground of the phones without affecting their normal operations of initiating or receiving calls or short messages. The 2.5 Million pilgrims are usually divided into groups and each group belongs to a Guide (Mutawif) that is responsible for the group from the time of arrival until the time of departure from the Kingdom of Saudi Arabia. The Guide takes care of the group residence, food and transportation. The developed system can be used to send location information to the group Guide in addition to sending it to the main pilgrimage authority server. The group Guide can also track any pilgrim and provide help when needed. The developed system could be used to track schoolchildren as well. The mobile phone of a child can be programmed to send location information to a server. The server can be accessed from the mobile of a parent or custodian of the child. Additionally, the system could be programmed to automatically give alarm if a child leaves a pre-specified region to his or her parents. Finally, if a user forgets or loses his mobile somewhere, then he can find its last detected location from the server, thus it would be easier to be found.

IV. CONCLUSION

The paper presents a system for pilgrim tracking and identification during Hajj in the Holy area using a mobile phone. The system consists of software that can be downloaded to the mobile phone of every pilgrim upon arrival to the Kingdom of Saudi Arabia. The mobile uses the Internet or SMS to send location information to a server managed by Hajj authority and to a server managed by the Guide of the group that the pilgrim is a member of. If the connection is lost, the mobile can store the location information in its memory until connection is restored. In that case the mobile sends the stored information to a server then clears its memory. The developed system provides an option for the pilgrims to request help in case of emergency. The location information is mapped onto a Google map or any geographical information system for ease

of localization and efficiency in providing needed help. A proof of concept experiment was implemented in the Holy area during the past pilgrimage season. The experiment has shown the viability of the proposed system for tracking pilgrims. For future work we plan to use mobile phones with NFC capability so that it can be used for identification as well as for tracking. Such a system will make a mobile phone be all what a pilgrim needs for his Hajj journey.

REFERENCES

- [1] Hajj and Umrah Statistics http://www.cdsi.gov.sa/pdf/Hajj_1435.pdf [Accessed: February 16, 2015]
- [2] Koshak, N.A., and Fouda, A., Analyzing pedestrian movement in Mataf using GPS and GIS to support space redesign. In Proceedings of the Ninth International Conference on Design and Decision Support Systems in Architecture and Urban Planning (The Netherlands, July 7-10, 2008)
- [3] Raising Number of Missing Pilgrims in Makkah. <http://archive.arabnews.com/?page=1§ion=0&article=87894&d=9&m=10&y=2006&pix=kingdom.jpg&category=Kingdom>, October 9, 2006 [Accessed: February 17, 2015]
- [4] Yamin, M. and Ades, Y. 2009. Crowd management with RFID and wireless technologies. In Proceedings of the 2009 First international Conference on Networks & Communications (December 27 - 29, 2009).NETCOM. IEEE Computer Society, Washington, DC, 439-442.
- [5] Luley, P., Almer, A., Seifert, C., Fritz, G., and Paletta, L. 2005. A Multi-Sensor system for mobile services with vision enhanced object and location awareness. In Proceedings of the Second IEEE International Workshop on Mobile Commerce and Services (July 18 - 19, 2005).WMCS. IEEE Computer Society, Washington, DC, 52-59.
- [6] Hasan, K.S., Rahman, M., Haque, A.L., Rahman, M.A., Rahman, T. and Rasheed, M.M., Cost effective GPS-GPRS based object tracking system. In Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 (Hong Kong, March 18 - 20, 2009)
- [7] M. Mohandes, "RFID-based System for Pilgrims Identification & Tracking", the Applied Computational Electromagnetics Society Journal (ACES), in press, 2009.
- [8] M. Mohandes, "Wireless Sensor Networks for Pilgrim Tracking", the 3rd Medina Monawara for Hajj Research, 2010.
- [9] Mohandes, Haleem, Abul-Hussain, and Balakrishnan, "Pilgrim Tracking using Wireless Sensor Network", Workshops of International Conference on Advanced Information Networking and Applications (WAINA 2011), Singapore, 22-25 March, 2011.
- [10] HTC - Products - HTC Touch Diamond2 - Specification. <http://www.htc.com/www/product/touchdiamond2/specification.html>, [Accessed: February 16, 2015]
- [11] Robusto, C. C., The Cosine-Haversine formula, The American Mathematical Monthly, Vol. 64, No. 1 (Jan., 1957), pp. 38-40. Internet:
- [12] <http://www.jstor.org/stable/2309088>, [Accessed: February 13, 2015]

ONLINE RECOGNITION SYSTEM FOR HANDWRITTEN ARABIC DIGITS

Mustafa Ali Abuzaraida

Computer Science Department
Faculty of Information Technology
Misurata University
Misurata, Libya
abuzaraida@umit.edu.ly

Akram M. Zeki

Kulliyyah of Information and Communication Technology
International Islamic University Malaysia
Kuala Lumpur, Malaysia
akramzeki@iium.edu.my

Ahmed M. Zeki

Department of Information Systems College of Information Technology University of Bahrain
Sakhir, Kingdom of Bahrain
amzeki@uob.edu.bh

Abstract— Nowadays, online text recognition systems are being given tremendous attention worldwide due to the fast growth of touch screen devices industry. Although, keyboards and mice devices have not become/are not applicable to be used by smaller devices, these reasons pushed researchers to focus on new techniques which are able to design this kind of systems. These online systems can deal with multiple types of texts such as alphabets, digits, and symbols. In this paper, an online system for recognizing handwritten Arabic digits has been presented. The paper illustrates four phases of the system in details which are: digits acquisition, preprocessing, features extraction, and recognition phase. The dataset of the system was collected by 100 writers using a touch screen laptop with 100 samples of each digit. The results of testing the proposed system showed a high accuracy rate with an average of 98 percentage.

Keywords: *Text Recognition, Online System, Handwritten Digits.*

I. INTRODUCTION

Text recognition field is considered as one of the major fields of the pattern recognition area which has been the subject of many researches in the past three decades [1].

Generally, offline text recognition approaches are designed to convert scanned scripts into a text documents. In contrast, online approaches capture the text by writing on touch screen devices or recording the movements of a stylus and convert the action into a text format.

Online recognition field has been gaining more interest lately due to the increasing of pen computing applications like tablet devices, digital notebooks, and advanced cellular phones [2]. Nowadays, these devices are commonly used worldwide that encouraged companies to improve their products to deal with multi languages. However, these devices can deal with many different languages spoken by billions of people around the world such as Latin, Chinese, Japanese, Indian, Korean, Arabic, and many others from textual or speech manner [3] [4].

From the literature in the text recognition field, it is noticeable that most of the research works were dedicated to offline approaches for Latin characters and other languages such Chinese. On the other hand, a few researches and studies

have been published to develop online approaches using new methods and algorithms in this area for texts in general and digits in particular [4].

Several studies have been published in this field during past decades. Most of the studies covered of solving segmentation problem and recognizing isolated characters while recognizing digits and mathematical symbols got less attention.

Arabic digits are commonly used by billions over the world. The shape of the digits (0,1,2,3,4,5,6,7,8,9) were originally designed by Arabic Mathematician scholars and upgraded by the Muslim scholar Al-Khwarizmi who invents the zero in the ninth century [5]. These styles were used in the western part of Arab world which located in North Africa and Alandalus "Spain" in the 10 century [6].

This paper is presenting an online system for recognizing handwritten Arabic digits. The system contains four main phases which are: text acquisition, preprocessing, feature extraction, and recognition phase.

The rest of this paper is organized as follows: Section II summarizes the architecture of the proposed system and each step of the system is explained. Section III presents the results

of testing the system while the conclusion and the summary of the paper is presented in section IV.

II. ARCHITECTURE OF THE PROPOSED SYSTEM

The proposed system followed the typical pattern recognition system architecture that contains four main phases which are: text acquisition, preprocessing, features extraction, and recognition phase [1] as shown in Figure 1. However, segmentation step is not included in the system and every handwritten digit is processed as one block. Segmentation free strategy can minimize the time process and can enhance the recognition accuracy rate [4]. Although, every phase of the system has one or more objectives in order to reach the system goal and also to enhance the overall recognition accuracy rate. The phases of the proposed system are explained as follows:

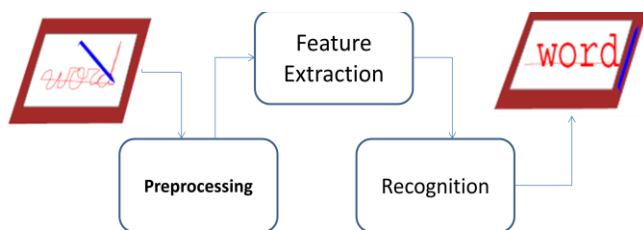


Figure 1. Typical phases of online Text Recognition System

A. Data Collection Stage

Data collection stage is the initial step of any pattern recognition system and aims to get raw data which is used later for training and testing manner. In this stage, every handwritten digit is captured by writing it on an interface device that records the handwritten digit in time stamped coordinates of pen trajectory (x, y) [3].

Here, for the purpose of collecting the training and testing datasets, 1.5 GHz core i3 Acer Tablet has been used to collect the dataset "same used in [7]". This touch screen computer can easily be used to acquire the handwritten Arabic digits with a simple way of normal writing on the touch screen using a special pen. The way of writing on this Tablet can minimize the noise and errors while recording on the Tablet surface.

For collecting handwritten Arabic digits, a platform was designed using Matlab environment with GUI interface. The writer can start writing the digit on the area of writing just after writing his/her writer identification number. The writer writes the digit that appears in provided image on the upper area of the acquiring data platform. After writing the digit, the writer should click on the next button to write the next digit. However, if the writer wants to rewrite the current digit before starting writing the next one, he/she should click on reset button to rewrite the digit again. Data collection by using this natural way of writing can provide closely resembles, smoothed, and filtered data collected from the computer Tablet. Figure 2 shows the data collection platform.

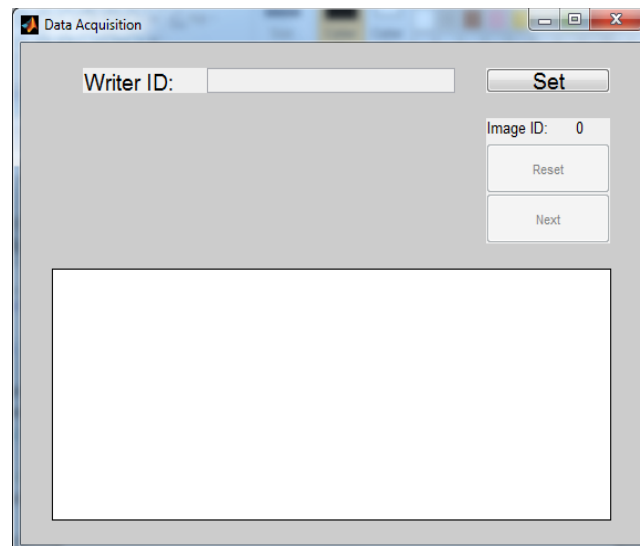


Figure 2. Data acquisition platform

B. Preprocessing Phase

Preprocessing phase structure is needed in both types of handwriting recognition systems[8]. However, due to the way of the text acquisition, the steps of preprocessing might be different. The noise in offline systems can occur because of many reasons such as the scanner quality, quality of the paper, and papers noise. Meanwhile, in online recognition systems, the noise can occur due to some other reasons like the form of sharp edges, non centered text, uneven sizes of text and missing points in text trajectories due to high speed of handwriting [3] [9].

Preprocessing phase in online handwriting recognition is performed to minimize the noise which may accrue in the handwritten text as mentioned earlier. This phase includes several multiple steps and every step does a specific function to filter the data set. Besides that, it can improve the overall recognition rate and considers as one of the essential phases of online handwriting recognition and most of the researchers have discussed its challenges for various texts from time to time [10].

Nevertheless, performing and many preprocessing steps in this phase may cause some problems in online handwritten recognition systems. For instance, delay may take place to overall time processing [4]. Also, it may affect and reduce the recognition accuracy rate by complicating the processing which can lead to omission of some important parts and features of the text [3].

Generally, data collection in online handwriting recognition stores the stylus movements on the writing surface. These movements are distributed at various positions on writing area of the acquisition platform and then joined from the first position (x₁, y₁) to the last (x_n, y_n) to present the appearance of drawn text. Although, the stylus movements consist of three actions which are: Pen Down, Pen Move and Pen Up actions.

The serial of points is collected when the writer presses, moves, lifts the stylus up consecutively. Pen Move function records the movements of the stylus on writing tablet from the writing start point (x_1, y_1) until the last point (x_n, y_n) where 'n' is the total number of points in the writing movements' list [11].

There are four steps which are included in preprocessing phase in the proposed system as follow:

i. *Digit Smoothing:*

In the proposed system, a smoothing technique is used to smoothen the handwritten curves called Loess filter. This filter is based on conducting the local regression of the curves points using weighted linear least squares and a second degree polynomial model.

In this technique, each smoothed value is determined locally by neighboring data points defined within the writing curve. The process is weighted and a regression weight function is defined for each data point contained within the writing curve [12].

ii. *Digit Simplification:*

Data point's simplification is the process of reducing the number of data points acquired by a digital device through removing the redundant points which could be inappropriate for pattern classification. This processing directly affects and enhances the recognizer performance. However, Douglas Peucker's algorithm [13] was adopted to simplify the acquired handwritten digit point sequence.

iii. *Digit Size Normalization:*

The size of the acquired handwritten digits depends on how the writer moves the stylus on writing area. The handwritten digits are generally written in different sizes when the pen is moved along the border of writing area that may cause some ambiguity in the next phases. Size normalization is a necessary step that should be performed in order to recognize any type of text. This can be achieved by converting the acquired handwritten digit with assumed fixed size format.

iv. *Centering of the Digit:*

After resizing the acquired handwritten digit, the current coordinates are needed to be shifted to the centering axis (x_0, y_0) to make sure that all handwritten digit points are in the equal formatting and all data are translated to the same spot relative to the origin.

C. *Features Extraction Phase*

For pattern recognition research field in general and text recognition approaches in particular, extracting an appropriate set of features and an efficient extraction method, have been considered as the most important factors for achieving high recognition performance [14].

In the feature extraction phase of the proposed system, each handwritten Arabic digit is described using a set of features

that distinguishes it from other handwritten digit in the dataset. The features of each handwritten digit represent in a ordered format called features vector. The features vector is then used in the next phase by the classifier to match the closest class using a classification criterion. In addition, the purpose of performing feature extraction phase is to realize that just a part of data points are equally important to the pattern recognition task.

In online recognition systems, the information about how the character has been written is found. Although, complex preprocessing steps cannot be performed in practical online systems such as Tablets and (personal digital assistant) PDAs since data is collected as the text is being written. Hence, taking advantage of the dynamic characteristics of the data is crucial such as the speed, angular velocity, and other features of this kind. These features remain available for processing as the character is written on the Tablet [14].

Choosing a proper type of features depends on the nature of the text, the type of the system processing which may be online or offline, and the texts types that can be handwritten or printed. However, feature types of recognizing any text can be categorized into three main types : structural features, statistical features, and global transformation [14].

Here, in this system, structural features are used to extract the handwritten Arabic digits features. While the system is not performing the segmentation part, the proposed system uses light amount of features which can help to avoid the complexity during the system excision.

Structural features are used to be the main features of the proposed system. They take the pen trajectory directions as the main feature representing handwriting movements. Freeman Chain code is used to create the direction matrix for each handwritten digit. Freeman Chain code [15] represents the pen movements directions by a numeric code consisting of 8 digits. These directions are listed from 1-8 to represent the eight main writing directions as illustrated in Figure 3.

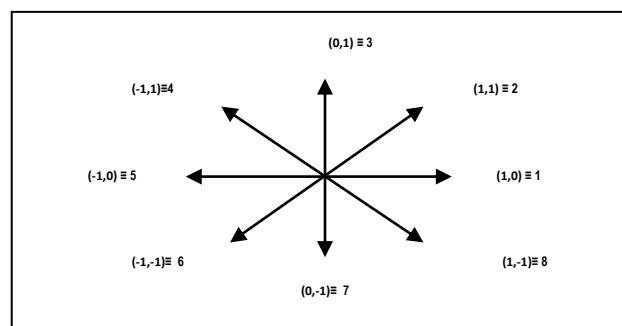


Figure 3. Freeman Chain Code

The process starts from the first point of the writing until the last point. The algorithm procedure is explained in the following steps:

- Read the points sequence S of the handwritten digit.

- Find the $d(x)$ and $d(y)$ values for the $S(x_2, y_2)$ point indexed by $S(x_1, y_1)$.
- Find the Freeman Chain Code for this pair from Table II.

TABLE II. FREEMAN CHAIN CODES

D(x)	D(y)	Code
0	+1	3
0	-1	7
-1	+1	4
-1	-1	6
+1	+1	2
+1	-1	8
-1	0	5
+1	0	1

- Make $S(x_2, y_2)$ as the first point and eliminate the previous point.
- Repeat Steps 2-4 for all the points.
- Record the Freeman Chain sequence to represent the handwritten digit movements' directions.

After completing the algorithm steps, a code for the directions of pen movements are stored. Each sequence is presenting a symbol of Arabic digit formatted in Freeman Chain Code. These sequences will be used in the next phase in order to distinguish and compare the handwritten digits by the classifier.

D. Recognition phase

In this study, matching algorithm called Global Alignment Algorithm (GAA) is used as recognition engine to recognize the Arabic digits. After conducting this phase, the system can classify the proper digit from the data set of the system [16].

Sequences Alignment or sequences comparison concenter is the heart of the bioinformatics field. It describes the way to arrange the DNA, RNA, or protein sequences by identifying the regions of similarity among them. Furthermore, it is used to conclude structural, functional, and evolutionary relationship between the matched sequences. Also, alignment algorithm finds the similarity level between query sequence and different database sequences. The algorithm is designed based on dynamic programming approach which divides the problem into smaller independent sub problems. It finds the alignment more quantitatively by assigning the matching scores [17].

In fact, the most well known and widely used methods for sequences alignments are: Local and Global Alignment Algorithms. Local Alignment Algorithm compares the sequences which are suspected to have similarity or even dissimilar sequences length to find the local regions with high level of similarity. On the other hand, it is very much appropriate to use Global Alignment Algorithm for comparing the closely related sequences which are of same length. Here, the alignment is carried out from the beginning until the end of the matched sequence to find out the best possible alignment

[16]. However, Global Alignment (Needleman-Wunsch algorithm) Algorithm is used in the proposed system as a classification/ classifying tool.

GAA was developed by Saul B. Needleman and Christian D. Wunsch in 1970 [16], which is basically a dynamic programming algorithm for sequence alignment. The dynamic programming can solve the original problem by dividing it into smaller independent sub-problems. The algorithm explains global sequence alignment for aligning nucleotide or protein sequences in general. However, these alignment techniques could be used in many different aspects of computer science approaches.

Basically, dynamic programming is used to find the optimal alignment of two sequences. It finds the alignment in a quantitative way by giving score values for matches and mismatches. The alignment is accurately obtained by searching the highest scores in the matrix [17].

For matching any two amino acid sequences, the algorithm is designed to find the highest score value of the sequences by building a two- dimensional matrix. Basically, the algorithm procedure is defined with three following steps.

- Assuming an initialization score matrix with the possible scores.
- Filling the matrix with maximum scores.
- For appropriate alignment, tracing back the previous maximum scores.

III. RESULTS OF TESTING THE SYSTEM

To test the proposed system, 50 writers were asked to write the Arabic digits. Every writer was asked to repeat writing every digit in his\her writing style to get 150 case for all the 10 digits in different writing style.

Using 20% of the digits dataset for testing and 80% for training, the system showed that most of the digits were recognized successfully. However, digits like 9 and 4 were recognized with some mismatching cases. This mismatching was accruing due to the similarity of the digits shapes.

IV. CONCLUSION AND SUMMARY

Recently, many countries start using smart tablets in classes to assist the students instead of writing on papers. Using these devices can make the learning much easier than using the typical learning way. These devices can support recognizing text software to recognize any kind of text. From this point of view, an online recognition system, to recognize Arabic texts, numbers, and symbols, is needed.

In this paper, a brief description of designing of recognition system for Arabic digits was highlighted. The main aim of this work was to open the research gate to this kind of research. On the other hand, this system can help to produce education software to recognize Arabic mathematical operations. This software can give more significance to/for

learning mathematical subjects in high schools or even in universities. This software can be of great significance in learning mathematical subjects at high school or even at university levels.

V. ACKNOWLEDGMENT

We would like to thank The Kulliyyah of Information and Communication Technology and the Research Management Center in the International Islamic University Malaysia for their supports.

REFERENCES

- [1] M. A. Abuzaraida, A. M. Zeki and A. M. Zeki, "Recognition Techniques for Online Arabic Handwriting Recognition Systems," In Proceeding of the International Conference on Advanced Computer Science Applications and Technologies (ACSAT2012), Kuala Lumpur, Malaysia, 2012.
- [2] Mustafa Ali Abuzaraida, Akram M Zeki and Ahmed M Zeki, "Online Recognition System for Handwritten Hindi Digits Based on Matching Alignment Algorithm," In Proceeding of the Third International Conference on Advanced Computer Science Applications and Technologies (ACSAT2014), Amman, Jordan, 2014.
- [3] Mustafa Ali Abuzaraida, Akram M. Zeki and Ahmed M. Zeki, "Problems of writing on digital surfaces in online handwriting recognition systems," In Proceeding of the Information and Communication Technology for the Muslim World (ICT4M), 2013 5th International Conference on, 2013, pp. 1-5.
- [4] M. A. Abuzaraida, A. M. Zeki and A. M. Zeki, "Segmentation Techniques for Online Arabic Handwriting Recognition: A survey," In Proceeding of the 3rd International Conference on Information and Communication Technology for the Moslem World: ICT Connecting Cultures, ICT4M 2010, Jakarta, Indonesia, 2010, pp. D37-D40.
- [5] R. Kaplan and E. Kaplan, *The Nothing that Is: A Natural History of Zero*: Oxford University Press, 1999.
- [6] Solomon Gandz, "The Origin of the Ghubār Numerals, or the Arabian Abacus and the Articuli." vol. 16, T. U. o. C. Press, Ed., ed: The University of Chicago Press, pp. 393-424, 1931.
- [7] Mustafa Ali Abuzaraida, Akram M Zeki and Ahmed M Zeki, "Online Database of Quranic Handwritten Words," *Journal of Theoretical & Applied Information Technology*, vol. 62, 2014.
- [8] Mustafa Ali Abuzaraida, Akram M Zeki, Ahmed M Zeki and Nor Farahidah Za'bah, "Online Recognition System for Handwritten Arabic Chemical Symbols," In Proceeding of the Computer and Communication Engineering (ICCCE), 2014 International Conference on, 2014, pp. 138-141.
- [9] M. A. Abuzaraida, A. M. Zeki and A. M. Zeki, "Difficulties and Challenges of Recognizing Arabic Text," in *Computer Applications: Theories and Applications*, ed Kuala Lumpur: IUM Press Malaysia, 2011.
- [10] N. Tagougui, M. Kherallah and A.M. Alimi, "Online Arabic handwriting recognition: a survey," *International Journal on Document Analysis and Recognition*, pp. 1-18, 2012.
- [11] Mai Al-Ammar, Reham Al-Majed and Hatim Aboalsamh, "Online Handwriting Recognition for the Arabic Letter Set," *Recent Researches in Communications and IT*, 2011.
- [12] Loader Clive, *Local Regression and Likelihood* vol. 47: springer New York, 1999.
- [13] Douglas David and Peucker Thomas, "Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or its Caricature," *Cartographica: The International Journal for Geographic Information and Geovisualization*, vol. 10, pp. 112-122, 1973.
- [14] M. A. Abuzaraida, Akram M Zeki and Ahmed M Zeki, "Feature Extraction Techniques of Online Handwriting Arabic Text Recognition," In Proceeding of the 5th International Conference on Information and Communication Technology for the Muslim World (ICT4M), 2013, pp. 1-7.
- [15] Freeman Herbert, "Computer Processing of Line-Drawing Images," *ACM Comput. Surv.*, vol. 6, pp. 57-97, 1974.
- [16] R Durbin, S Wddy, A Korgh and G Mitchison, *Biological sequence analysis: probabilistic models of proteins and nucleic acids*: Cambridge University Press, 1998.
- [17] Neil C. Jones and Pavel A. Pevzner, *An Introduction to Bioinformatics Algorithms*, illustrated ed. Cambridge, Massachusetts London, England: Massachusetts Institute of Technology Press, 2004.

Hierarchical Singular Value Decomposition for Halftone Images

Roumen Kountchev
Department of RCTV
Technical University of Sofia
Sofia, Bulgaria
rkountch@tu-sofia.bg

Roumiana Kountcheva
TK Engineering
Sofia
Bulgaria
kountcheva_4@yahoo.com

Abstract - This work is devoted to one new approach for decomposition of images represented by matrices of size $2^n \times 2^n$, based on the multiple application of the Singular Value Decomposition (SVD) over image blocks of relatively small size (2×2), obtained after division of the original image matrix. The new decomposition, called Hierarchical SVD, has tree structure of the kind binary tree of n hierarchical levels. Its basic advantages over the famous SVD are: the reduced computational complexity, the opportunity for parallel and recursive processing of the image blocks, based on relatively simple algebraic relations, the high concentration of the image energy in the first decomposition components, and the ability to accelerate the calculations through cutting-off the tree branches in the decomposition levels, where the corresponding eigen values are very small. The HSVD algorithm is generalized for images of unspecified size. The new decomposition opens numerous opportunities for fast image processing in various application areas: image compression, filtration, segmentation, merging, digital watermarking, extraction of minimum number of features sufficient for the objects recognition, etc.

Keywords - Singular Value Decomposition (SVD), block SVD, Hierarchical SVD, binary tree, computational complexity.

I. INTRODUCTION

The SVD decomposition had significant influence on the processing and analysis of digital images used in computer vision systems. This decomposition was the target of significant number of investigations, presented in scientific monographs [1-6] and papers [7-12].

The SVD has the following basic features: 1) it is an optimum decomposition, because it concentrates maximum part of the image energy in a minimum number of components; 2) the image, restored after the reduction of the low-energy components has minimum mean square error. One of the basic problems, which restrict the practical use of the famous SVD, is its high computational complexity, which grows together with the size of the image matrix. Several approaches are offered to overcome this problem. The first approach is based on the SVD calculation through iterative methods, which do not demand to define the characteristic polynomial of the matrix. In this case the SVD is executed in two stages: in the first, the matrix is transformed into triangular form through the QR decomposition, and then - into bi-diagonal through Householder's transforms [13]; in the second stage on the bi-diagonal matrix is applied an iterative algorithm, whose iterations stop when the needed accuracy is obtained. Such is,

for example, the iterative method of Jacobi [3, 6, 25], in accordance with which to calculate the SVD for a bi-diagonal matrix, is needed to execute a sequence of orthogonal transforms with matrices, which differ from the singular matrix in the elements of the rotation matrix of size 2×2 only. The second approach is based on the relation between the SVD and the Principal Component Analysis (PCA). It could be implemented through neural networks [14] of the kind generalized Hebbian or multilayer perceptron networks, which use iterative learning algorithms. One more approach is based on the algorithm, known as Sequential KL/SVD [15]. Its basic idea is given in brief as follows: the image matrix is divided into blocks of small size, on which is applied SVD, based on the QR decomposition [6]. The SVD is initially calculated for the first block, and then iterative SVD calculation is executed for each block, using the transform matrix, already obtained for the preceding block (update procedure). In the iteration process the SVD components, which have very small values, are eliminated.

II. RELATED WORK

Several methods had already been developed, aimed at the enhancement of the SVD calculation [16-19]. The first, called

Randomized SVD [16, 17], is based on the algorithm in accordance with which, are randomly selected some rows (or columns) of the transform matrix. After scaling, they build a small matrix, for which is calculated the SVD, which is then used as an approximation of the original matrix. In [18] is offered the QUIC-SVD algorithm, which is suitable for matrices of very large size. Using this algorithm is achieved fast sample-based SVD approximation with automatic relative error control. This algorithm also uses a sampling mechanism, called “the cosine tree”, to achieve best-rank approximation. The experimental investigation on the QUIC-SVD, given in [19], offers better results than these, obtained with MATLAB SVD and Tygert SVD [17]. The speedup achieved is 6-7 times higher compared to that of the exact SVD, but it depends on the selected value for the parameter δ which defines the higher limit of the approximation error with a probability of size $(1-\delta)$.

Significant number of SVD-based methods had been developed, aimed at the image compression efficiency enhancement [20-24]. The method, called Multiresolution SVD [20], comprises 3 steps: 1) image transform through 9/7 bi-orthogonal wavelets of 2 levels; 2) decomposition of the transformed image through SVD executed on blocks of size 2×2 up to level six, and 3) execution of the SPIHT and gzip algorithms. In [21] is offered a hybrid KLT-SVD algorithm for efficient image compression. The K-SVD [22] for facial image compression is a generalization of the K-means clusterization method and is applied in the iterative learning of over-complete sparse coding dictionaries. In correspondence with the combined compression algorithm presented in [23], the SVD is executed individually for each of the color components R, G, B, segregated from the image stored in the JPEG file format. In [24] is introduced the Higher-Order SVD (HOSVD), which is an extension of the SVD matrix to tensors with application in the data compression. In [26, 27] are presented some parallel hardware implementations of the SVD for symmetrical matrices, based on the Jacobi’s method.

In this work is offered one new approach for hierarchical image decomposition, based on the multiple SVD execution on blocks of small size. This decomposition, called here the “Hierarchical SVD” (HSVD), has a tree structure of the kind binary or 3-nodes tree (full or truncated). The SVD calculation for blocks of size 2×2 is based on the adaptive KLT [28]. The HSVD algorithm [29, 30] is aimed at the achievement of decomposition with high computational efficiency, which is also suitable for parallel recursive processing with relatively simple algebraic operations, and permits calculation speedup through cutting-off the branches with very small eigenvalues.

The paper comprises the following sections: SVD calculation for a matrix of size 2×2 ; representation of the hierarchical SVD for a matrix of size $2^n \times 2^n$; evaluation of the computational complexity of the hierarchical SVD of size $2^n \times 2^n$; representation of the HSVD algorithm through tree-like structure, and conclusions.

III. CALCULATION OF SVD WITH A MATRIX OF SIZE 2×2

A. General case: SVD execution on image of size $N \times N$

In the general case, the decomposition of the square image $[X(N)]$, represented by a matrix of size $N \times N$ is based on the direct SVD, defined by the relation below [10, 11]:

$$[X(N)] = [U(N)][\Lambda(N)]^{1/2}[V(N)]^t = \sum_{s=1}^N \sqrt{\lambda_s} \vec{U}_s \vec{V}_s^t \quad (1)$$

The inverse SVD is respectively represented as:

$$[\Lambda(N)]^{1/2} = [U(N)]^t [X(N)] [V(N)]. \quad (2)$$

In the equations above, the terms $[U(N)] = [\vec{U}_1, \vec{U}_2, \dots, \vec{U}_N]$ and $[V(N)] = [\vec{V}_1, \vec{V}_2, \dots, \vec{V}_N]$ are matrices, composed by the vectors \vec{U}_s and \vec{V}_s for $s=1, 2, \dots, N$. Here \vec{U}_s are the eigen vectors of the matrix $[Y(N)] = [X(N)]^t [X(N)]$ (left-singular vectors of $[X(N)]$), and \vec{V}_s are the eigen vectors of the matrix $[Z(N)] = [X(N)][X(N)]^t$ (right-singular vectors of $[X(N)]$), for which:

$$[Y(N)]\vec{U}_s = \lambda_s \vec{U}_s, \quad (3)$$

$$[Z(N)]\vec{V}_s = \lambda_s \vec{V}_s. \quad (4)$$

$[\Lambda(N)] = \text{diag}[\lambda_1, \lambda_2, \dots, \lambda_N]$ is a diagonal matrix, composed of the eigenvalues λ_s of both matrices $[Y(N)]$ and $[Z(N)]$, which are same.

From (1) it follows that for the description of a matrix of size $N \times N$ are needed $N \times (2N+1)$ parameters in total, i.e., in the general case the SVD is an of over-complete decomposition.

B. Particular case: SVD for one image block of size 2×2

The direct SVD for the square block $[X]$ of size 2×2 ($N=2$) is represented by the relation:

$$[X] = \begin{bmatrix} a & b \\ c & d \end{bmatrix} = [U][\Lambda]^{1/2}[V]^t = \sqrt{\lambda_1} \vec{U}_1 \vec{V}_1^t + \sqrt{\lambda_2} \vec{U}_2 \vec{V}_2^t = \sum_{s=1}^2 \sqrt{\lambda_s} \vec{U}_s \vec{V}_s^t, \quad (5)$$

where a, b, c, d are pixels; λ_1, λ_2 - common eigen values of the symmetrical matrices $[Y]$ and $[Z]$:

$$[Y] = [X]^t [X] = \begin{bmatrix} a & c \\ b & d \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} (a^2 + c^2) & (ab + cd) \\ (ab + cd) & (b^2 + d^2) \end{bmatrix}; \quad (6)$$

$$[Z] = [X][X]^t = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} a & c \\ b & d \end{bmatrix} = \begin{bmatrix} (a^2 + b^2) & (ac + bd) \\ (ac + bd) & (c^2 + d^2) \end{bmatrix}. \quad (7)$$

\vec{U}_1 and \vec{U}_2 are the eigenvectors of the matrix $[Y]$, for which: $[Y]\vec{U}_s = \lambda_s \vec{U}_s, s = 1, 2;$

\vec{V}_1 and \vec{V}_2 are the eigenvectors of the matrix $[Z]$, for which: $[Z]\vec{V}_s = \lambda_s \vec{V}_s, s = 1, 2.$

$[U]=[\vec{U}_1, \vec{U}_2]$ and $[V]^t = \begin{bmatrix} \vec{V}_1^t \\ \vec{V}_2^t \end{bmatrix}$ - matrices, composed of the eigenvectors \vec{U}_s and \vec{V}_s .

C. Calculation of the eigenvalues and vectors of the symmetrical matrix of size 2x2

Let for $N=2$ the corresponding matrix $[G] = \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix}$ is symmetrical in respect of its main diagonal. Then here could be assumed the simplified symbols: $g_{11}=g_1, g_{22}=g_2, g_{12}=g_{21}=g_3$. The eigenvalues λ_1, λ_2 of the matrix $[G]$ are the solution of the characteristic equation:

$$\det[G - \lambda I] = \lambda^2 - (g_1 + g_2)\lambda + (g_1g_2 - g_3^2) = 0. \quad (8)$$

Since the matrix $[G]$ is symmetrical, its eigenvalues are real numbers:

$$\lambda_1 = \frac{1}{2} \left[(g_1 + g_2) + \sqrt{(g_1 - g_2)^2 + 4g_3^2} \right], \quad (9)$$

$$\lambda_2 = \frac{1}{2} \left[(g_1 + g_2) - \sqrt{(g_1 - g_2)^2 + 4g_3^2} \right].$$

The eigenvectors $\vec{\Phi}_s$ of the matrix $[G]$ for $s=1, 2$ are the solutions of the system of equations below:

$$(g_1 - \lambda_s)\Phi_{1s} + g_3\Phi_{2s} = 0,$$

$$g_1\Phi_{1s} + (g_2 - \lambda_s)\Phi_{2s} = 0, \quad (10)$$

$$\Phi_{1s}^2 + \Phi_{2s}^2 = 1.$$

The eigenvector $\vec{\Phi}_s = [\Phi_{1s}, \Phi_{2s}]^T$, which corresponds to the eigenvalue λ_s , is:

- For $s=1$

$$\vec{\Phi}_1 = \frac{1}{\sqrt{2[(g_1 - g_2)^2 + 4g_3^2] + (g_1 - g_2)\sqrt{(g_1 - g_2)^2 + 4g_3^2}}} \times$$

$$[(g_1 - g_2) + \sqrt{(g_1 - g_2)^2 + 4g_3^2}, 2g_3]^t = \frac{1}{\sqrt{2\gamma(\gamma + \alpha)}} [\alpha + \gamma, \beta]^t, \quad (11)$$

- For $s=2$

$$\vec{\Phi}_2 = \frac{1}{\sqrt{2[(g_1 - g_2)^2 + 4g_3^2] - (g_1 - g_2)\sqrt{(g_1 - g_2)^2 + 4g_3^2}}} \times$$

$$[(g_1 - g_2) - \sqrt{(g_1 - g_2)^2 + 4g_3^2}, 2g_3]^t = \frac{1}{\sqrt{2\gamma(\gamma - \alpha)}} [\alpha - \gamma, \beta]^t, \quad (12)$$

$$\alpha = g_1 - g_2, \gamma = \sqrt{\alpha^2 + \beta^2} = \sqrt{(g_1 - g_2)^2 + 4g_3^2} \text{ and } \beta = 2g_3$$

The matrix $[\Phi]$ consists of the eigenvectors $\vec{\Phi}_1 = [\Phi_{11}, \Phi_{21}]^t$ and $\vec{\Phi}_2 = [\Phi_{12}, \Phi_{22}]^t$:

$$[\Phi] = \begin{bmatrix} \vec{\Phi}_1^t \\ \vec{\Phi}_2^t \end{bmatrix} = \begin{bmatrix} \Phi_{11} & \Phi_{21} \\ \Phi_{12} & \Phi_{22} \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} \frac{\alpha + \gamma}{\sqrt{\gamma^2 + \alpha\gamma}} & \frac{\beta}{\sqrt{\gamma^2 + \alpha\gamma}} \\ \frac{\alpha - \gamma}{\sqrt{\gamma^2 - \alpha\gamma}} & \frac{\beta}{\sqrt{\gamma^2 - \alpha\gamma}} \end{bmatrix}. \quad (13)$$

Then the corresponding transposed matrix is:

$$[\Phi]^t = [\vec{\Phi}_1, \vec{\Phi}_2] = \begin{bmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{bmatrix} =$$

$$= \frac{1}{\sqrt{2}} \begin{bmatrix} \frac{\alpha + \gamma}{\sqrt{\gamma^2 + \alpha\gamma}} & \frac{\alpha - \gamma}{\sqrt{\gamma^2 - \alpha\gamma}} \\ \frac{\beta}{\sqrt{\gamma^2 + \alpha\gamma}} & \frac{\beta}{\sqrt{\gamma^2 - \alpha\gamma}} \end{bmatrix} \quad (14)$$

The elements Φ_{ij} of the matrix $[\Phi]$ could be represented as a function of the angle θ , on which the coordinate system, defined by the vectors $\vec{\Phi}_1$ and $\vec{\Phi}_2$ is rotated in respect to the original coordinate system. In this case:

$$[\Phi(\theta)] = \begin{bmatrix} \Phi_{11}(\theta) & \Phi_{21}(\theta) \\ \Phi_{12}(\theta) & \Phi_{22}(\theta) \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}, \quad (15)$$

$$\theta = \arctg\left(\frac{\Phi_{21}(\theta)}{\Phi_{11}(\theta)}\right) = \arctg\left(\frac{\beta}{\alpha + \gamma}\right) =$$

where

$$= \arctg\left(\frac{2g_3}{(g_2 - g_1) + \sqrt{(g_1 - g_2)^2 + 4g_3^2}}\right).$$

The elements Φ_{ij} of the matrix $[\Phi(\theta)]$ are:

$$\cos\theta = \cos\left[\arctg\left(\frac{\beta}{\alpha + \gamma}\right)\right] =$$

$$= \frac{\alpha + \gamma}{\sqrt{2\gamma(\gamma + \alpha)}} = \frac{\beta}{\sqrt{2\gamma(\gamma - \alpha)}}, \quad (16)$$

$$\sin\theta = \sin\left[\arctg\left(\frac{\beta}{\alpha + \gamma}\right)\right] =$$

$$= \frac{\beta}{\sqrt{2\gamma(\gamma + \alpha)}} = -\frac{\alpha - \gamma}{\sqrt{2\gamma(\gamma - \alpha)}}. \quad (17)$$

Since $\tan 2\theta = (2\tan\theta)/(1 - \tan^2\theta)$, the angle θ is defined as:

$$\theta = \frac{1}{2} \arctg\left(\frac{\beta}{\alpha}\right) = \frac{1}{2} \arctg\left(\frac{2g_3}{g_1 - g_2}\right). \quad (18)$$

From (15) it follows, that:

$$[\Phi(\theta)] = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} = \frac{1}{\sqrt{2\gamma(\alpha + \gamma)}} \begin{bmatrix} 1 & \frac{\beta}{\alpha + \gamma} \\ -\frac{\beta}{\alpha + \gamma} & 1 \end{bmatrix}. \quad (19)$$

In this case the eigenvectors are correspondingly [6]:

$$\vec{\Phi}_1 = [\cos\theta, \sin\theta]^t \text{ and } \vec{\Phi}_2 = [-\sin\theta, \cos\theta]^t, \quad (20)$$

where the angle θ is defined by (18).

D. Calculation of the eigenvalues and the eigenvectors of matrices [Y] and [Z]

The characteristic equation of the matrices [Y] and [Z], defined in accordance with (8), is:

$$\lambda^2 - (a^2 + b^2 + c^2 + d^2)\lambda + (ad + bc)^2 = 0. \quad \alpha = g_1 - g_2, \quad \beta = 2g_3, \quad \gamma = \sqrt{\alpha^2 + \beta^2} = \sqrt{(g_1 - g_2)^2 + 4g_3^2} \quad (21)$$

On the basis of (20), (11) and (12) are calculated the values of λ_s , \vec{U}_s and \vec{V}_s for $s = 1, 2$:

$$\lambda_{1,2} = \frac{1}{2} \left[(a^2 + b^2 + c^2 + d^2) \pm \sqrt{(a^2 + c^2 - b^2 - d^2)^2 + 4(ab + cd)^2} \right] = \quad (22)$$

$$= \frac{1}{2} (\omega \pm \sqrt{v^2 + 4\eta^2}) = \frac{1}{2} (\omega \pm A),$$

$$\vec{U}_1 = \frac{1}{\sqrt{2(v^2 + 4\eta^2 + v\sqrt{v^2 + 4\eta^2})}} [v + \sqrt{v^2 + 4\eta^2}, 2\eta]^t = \quad (23)$$

$$= \frac{1}{\sqrt{2A(A+v)}} [v+A, 2\eta]^t,$$

$$\vec{U}_2 = \frac{1}{\sqrt{2(v^2 + 4\eta^2 - v\sqrt{v^2 + 4\eta^2})}} [v - \sqrt{v^2 + 4\eta^2}, 2\eta]^t = \quad (24)$$

$$= \frac{1}{\sqrt{2A(A-v)}} [v-A, 2\eta]^t,$$

$$\vec{V}_1 = \frac{1}{\sqrt{2(\mu^2 + 4\delta^2 + \mu\sqrt{\mu^2 + 4\delta^2})}} [\mu + \sqrt{\mu^2 + 4\delta^2}, 2\delta]^t = \quad (25)$$

$$= \frac{1}{\sqrt{2B(B+\mu)}} [\mu+B, 2\delta]^t,$$

$$\vec{V}_2 = \frac{1}{\sqrt{2(\mu^2 + 4\delta^2 - \mu\sqrt{\mu^2 + 4\delta^2})}} [\mu - \sqrt{\mu^2 + 4\delta^2}, 2\delta]^t = \quad (26)$$

$$= \frac{1}{\sqrt{2B(B-\mu)}} [\mu-B, 2\delta]^t,$$

where:

$$\omega = a^2 + b^2 + c^2 + d^2, \quad v = a^2 + c^2 - b^2 - d^2, \quad \mu = a^2 + b^2 - c^2 - d^2, \quad (27)$$

$$\eta = ab + cd, \quad \delta = ac + bd, \quad A = \sqrt{v^2 + 4\eta^2}, \quad B = \sqrt{\mu^2 + 4\delta^2}. \quad (28)$$

The direct SVD for a matrix of size 2×2 could be represented by the relation:

$$[X] = \begin{bmatrix} a & b \\ c & d \end{bmatrix} = [U][\Lambda]^{1/2}[V]^t = \quad (29)$$

$$= \sqrt{\lambda_1}[T_1] + \sqrt{\lambda_2}[T_2] = \sum_{s=1}^2 \sqrt{\lambda_s}[T_s]$$

where $[U] = [\vec{U}_1, \vec{U}_2]$, $[\Lambda] = \text{diag}[\lambda_1, \lambda_2]$, $[V] = [\vec{V}_1, \vec{V}_2]$.

The eigen images of the matrix [X] are the matrices $[T_1]$ and $[T_2]$, defined by the relations:

$$[T_1] = \vec{U}_1 \vec{V}_1^t = \frac{1}{\sqrt{4AB(A+v)(B+\mu)}} \begin{bmatrix} (v+A)(\mu+B) & 2(v+A)\delta \\ 2(\mu+B)\eta & 4\eta\delta \end{bmatrix}, \quad (30)$$

$$[T_2] = \vec{U}_2 \vec{V}_2^t = \frac{1}{\sqrt{4AB(A-v)(B-\mu)}} \begin{bmatrix} (v-A)(\mu-B) & 2(v-A)\delta \\ 2(\mu-B)\eta & 4\eta\delta \end{bmatrix} \quad (31)$$

If the vectors \vec{U}_s and \vec{V}_s are defined in accordance with (20), the eigen images are:

$$[T_1] = \vec{U}_1 \vec{V}_1^t = \begin{bmatrix} \cos\theta_1 \\ \sin\theta_1 \end{bmatrix} [\cos\theta_2, \sin\theta_2] = \quad (32)$$

$$\begin{bmatrix} \cos\theta_1 \cos\theta_2 & \cos\theta_1 \sin\theta_2 \\ \sin\theta_1 \cos\theta_2 & \sin\theta_1 \sin\theta_2 \end{bmatrix},$$

$$[T_2] = \vec{U}_2 \vec{V}_2^t = \begin{bmatrix} -\sin\theta_1 \\ \cos\theta_1 \end{bmatrix} [-\sin\theta_2, \cos\theta_2] = \quad (33)$$

$$= \begin{bmatrix} \sin\theta_1 \sin\theta_2 & -\sin\theta_1 \cos\theta_2 \\ -\cos\theta_1 \sin\theta_2 & \cos\theta_1 \cos\theta_2 \end{bmatrix},$$

$$\theta_1 = \frac{1}{2} \arctg\left(\frac{2\eta}{v}\right), \quad \theta_2 = \frac{1}{2} \arctg\left(\frac{2\delta}{\mu}\right). \quad (34)$$

The inverse SVD for a matrix of size 2×2 is defined by the relation:

$$[\Lambda]^{1/2} = \begin{bmatrix} \lambda_1^{1/2} & 0 \\ 0 & \lambda_2^{1/2} \end{bmatrix} = [U]^t [X] [V], \quad (35)$$

where

$$[U]^t = \begin{bmatrix} \vec{U}_1^t \\ \vec{U}_2^t \end{bmatrix} = \begin{bmatrix} U_{11} & U_{21} \\ U_{12} & U_{22} \end{bmatrix} = \quad (36)$$

$$\frac{1}{\sqrt{2}} \begin{bmatrix} \frac{v+A}{\sqrt{A(A+v)}} & \frac{2\eta}{\sqrt{A(A+v)}} \\ \frac{v-A}{\sqrt{A(A-v)}} & \frac{2\eta}{\sqrt{A(A-v)}} \end{bmatrix} = \begin{bmatrix} \cos\theta_1 & \sin\theta_1 \\ -\sin\theta_1 & \cos\theta_1 \end{bmatrix},$$

$$[V] = [\vec{V}_1, \vec{V}_2] = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} \frac{\mu+B}{\sqrt{B(B+\mu)}} & \frac{\mu-B}{\sqrt{B(B-\mu)}} \\ \frac{2\delta}{\sqrt{B(B+\mu)}} & \frac{2\delta}{\sqrt{B(B-\mu)}} \end{bmatrix} = \quad (37)$$

$$= \begin{bmatrix} \cos\theta_2 & -\sin\theta_2 \\ \sin\theta_2 & \cos\theta_2 \end{bmatrix}.$$

The couple Direct/Inverse SVD could be then represented as follows:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \frac{I}{2\sqrt{2}} \left\{ \begin{aligned} &\sqrt{\frac{\omega + A}{AB(A+v)(B+\mu)}} \begin{bmatrix} v+A \\ 2\eta \end{bmatrix} [\mu+B, 2\delta] + \\ &+ \sqrt{\frac{\omega - A}{AB(A-v)(B-\mu)}} \begin{bmatrix} v-A \\ 2\eta \end{bmatrix} [\mu-B, 2\delta] \end{aligned} \right\} = \quad (38)$$

$$= \frac{I}{2\sqrt{2}} \left\{ \begin{aligned} &\sqrt{\frac{\omega + A}{AB(A+v)(B+\mu)}} \begin{bmatrix} (v+A)(\mu+B) & 2(v+A)\delta \\ 2(\mu+B)\eta & 4\eta\delta \end{bmatrix} + \\ &+ \sqrt{\frac{\omega - A}{AB(A-v)(B-\mu)}} \begin{bmatrix} (v-A)(\mu-B) & 2(v-A)\delta \\ 2(\mu-B)\eta & 4\eta\delta \end{bmatrix} \end{aligned} \right\}$$

or

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \sigma_I \begin{bmatrix} \cos\theta_1 \cos\theta_2 & \cos\theta_1 \sin\theta_2 \\ \sin\theta_1 \cos\theta_2 & \sin\theta_1 \sin\theta_2 \end{bmatrix} + \quad (39)$$

$$+ \sigma_2 \begin{bmatrix} \sin\theta_1 \sin\theta_2 & -\sin\theta_1 \cos\theta_2 \\ -\cos\theta_1 \sin\theta_2 & \cos\theta_1 \cos\theta_2 \end{bmatrix}$$

$$\begin{bmatrix} \sigma_I & 0 \\ 0 & \sigma_2 \end{bmatrix} = \frac{I}{2} \begin{bmatrix} \frac{v+A}{\sqrt{A(A+v)}} & \frac{2\eta}{\sqrt{A(A+v)}} \\ \frac{v-A}{\sqrt{A(A-v)}} & \frac{2\eta}{\sqrt{A(A-v)}} \end{bmatrix} \times \quad (40)$$

$$\times \begin{bmatrix} a & b \\ c & d \end{bmatrix} \times \begin{bmatrix} \frac{\mu+B}{\sqrt{B(B+\mu)}} & \frac{\mu-B}{\sqrt{B(B-\mu)}} \\ \frac{2\delta}{\sqrt{B(B+\mu)}} & \frac{2\delta}{\sqrt{B(B-\mu)}} \end{bmatrix}$$

$$\begin{bmatrix} \sigma_I & 0 \\ 0 & \sigma_2 \end{bmatrix} = \begin{bmatrix} \cos\theta_1 & \sin\theta_1 \\ -\sin\theta_1 & \cos\theta_1 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} \cos\theta_2 & -\sin\theta_2 \\ \sin\theta_2 & \cos\theta_2 \end{bmatrix} \quad (41)$$

$$\text{where } \sigma_I = \sqrt{\lambda_1} = \sqrt{\frac{\omega+A}{2}} \text{ and } \sigma_2 = \sqrt{\lambda_2} = \sqrt{\frac{\omega-A}{2}}. \quad (42)$$

Check: if $a=b=c=d$, then $\omega = 4a^2$, $\mu = v = 0$, $\delta = \eta = 2a^2$, $A = B = \sqrt{4^2 a^4} = 4a^2$, $\theta_1 = \theta_2 = \pi/4$.

In this case the couple Direct/Inverse SVD is correspondingly:

$$\begin{bmatrix} a & a \\ a & a \end{bmatrix} = \sqrt{\frac{4a^2 + 4a^2}{2}} \begin{bmatrix} (\sqrt{2}/2)^2 & (\sqrt{2}/2)^2 \\ (\sqrt{2}/2)^2 & (\sqrt{2}/2)^2 \end{bmatrix} + \quad (43)$$

$$+ \sqrt{\frac{4a^2 - 4a^2}{2}} \begin{bmatrix} (\sqrt{2}/2)^2 & -(\sqrt{2}/2)^2 \\ -(\sqrt{2}/2)^2 & (\sqrt{2}/2)^2 \end{bmatrix} =$$

$$= 2a \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = a \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 2a & 0 \\ 0 & 0 \end{bmatrix} = \frac{I}{2} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} a & a \\ a & a \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} = \quad (44)$$

$$= \frac{I}{2} \begin{bmatrix} 2a & 2a \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} = \frac{I}{2} \begin{bmatrix} 4a & 0 \\ 0 & 0 \end{bmatrix}$$

The equations above confirm the correctness of (39) and (41) for the SVD of size 2×2 , executed for the image matrix of same size and with constant brightness of the image pixels. From (39) it follows, that that for the representation of the matrix $[X]$ of size 2×2 through SVD are needed four parameters altogether: σ_1 , σ_2 , θ_1 and θ_2 , calculated on the basis of (22) and (32). Hence, the SVD of size 2×2 (SVD $_{2 \times 2}$), defined in accordance with (39), is not over-complete.

E. Energy distribution in the SVD components for a matrix of size 2×2

The energy of the matrix $[X] = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ (or of its quadratic Euclidean norm) is defined by the sum of the squares of its elements:

$$E_X = \sum_{i=1}^2 \sum_{j=1}^2 x_{i,j}^2 = a^2 + b^2 + c^2 + d^2 = \omega. \quad (43)$$

In correspondence with (5) and (39) the matrix $[X]$ is represented as the sum of two components:

$$[X] = \sigma_I [T_1] + \sigma_2 [T_2] = [C_1] + [C_2]. \quad (44)$$

$$[C_1] = \begin{bmatrix} c_{11}(I) & c_{12}(I) \\ c_{13}(I) & c_{14}(I) \end{bmatrix} = \quad (45)$$

$$= \sigma_I \begin{bmatrix} \cos\theta_1 \cos\theta_2 & \cos\theta_1 \sin\theta_2 \\ \sin\theta_1 \cos\theta_2 & \sin\theta_1 \sin\theta_2 \end{bmatrix},$$

$$[C_2] = \begin{bmatrix} c_{11}(2) & c_{12}(2) \\ c_{13}(2) & c_{14}(2) \end{bmatrix} = \quad (46)$$

$$= \sigma_2 \begin{bmatrix} \sin\theta_1 \sin\theta_2 & -\sin\theta_1 \cos\theta_2 \\ -\cos\theta_1 \sin\theta_2 & \cos\theta_1 \cos\theta_2 \end{bmatrix}$$

C_1 and C_2 are the eigen images of the matrix $[X]$.

The energy of each eigen image $[C_1]$, $[C_2]$ is respectively:

$$E_{C_1} = \sum_{i=1}^2 \sum_{j=1}^2 c_{i,j}^2(I) = \sigma_I^2 = \frac{\omega + A}{2}, \quad (47)$$

$$E_{C_2} = \sum_{i=1}^2 \sum_{j=1}^2 c_{i,j}^2(2) = \sigma_2^2 = \frac{\omega - A}{2}. \quad (48)$$

From the Parseval's theorem for energy preservation, ($E_X = E_{C_1} + E_{C_2}$) and from (47) and (48) it follows, that $E_{C_1} \gg E_{C_2}$, i.e., the energy E_X of the matrix $[X]$ is concentrated mainly in the first SVD component. The concentration degree is defined by the relation:

$$\xi = \frac{E_{C_1}}{E_{C_1} + E_{C_2}} = \frac{\sigma_I^2}{\sigma_I^2 + \sigma_2^2} = \frac{\omega + A}{2\omega}. \quad (49)$$

In particular, for the case, when the matrix $[X]$ is with equal values of the elements ($x_{i,j} = a$), from (39), (47), (48) and (49) is obtained $E_X = E_{C_1} = 4a^2$, $E_{C_2} = 0$ and $\xi = 1$. Hence, the total energy of the matrix $[X]$ is concentrated in the first SVD component only.

IV. HIERARCHICAL SVD FOR A MATRIX OF SIZE $2^N \times 2^N$

The hierarchical SVD (HSVD) for the image matrix $[X(N)]$ of size $2^n \times 2^n$ pixels ($N=2^n$) is implemented through multiple execution of the n -levels SVD on image blocks (sub-matrices) of size 2×2 . Let the matrix $[X(4)]$ is of size $2^2 \times 2^2$ ($N=2^2=4$). In this case the number of hierarchical levels is $n=2$.

In the first HSVD level ($r=1$), the matrix $[X(4)]$ is divided into 4 sub-matrices of size 2×2 , as it is shown in the left part of Fig. 1. The elements of the sub-matrices are as follows:

$$[X(4)] = \begin{bmatrix} [X_1(2)] & [X_2(2)] \\ [X_3(2)] & [X_4(2)] \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix} & \begin{bmatrix} a_2 & b_2 \\ c_2 & d_2 \end{bmatrix} \\ \begin{bmatrix} a_3 & b_3 \\ c_3 & d_3 \end{bmatrix} & \begin{bmatrix} a_4 & b_4 \\ c_4 & d_4 \end{bmatrix} \end{bmatrix}. \quad (50)$$

On each sub-matrix $[X_k(2)]$ of size 2×2 ($k=1,2,3,4$) is applied $SVD_{2 \times 2}$, in accordance with (39). As a result, it is decomposed into 2 components:

$$[X_k(2)] = \sigma_{1,k} [T_{1,k}(2)] + \sigma_{2,k} [T_{2,k}(2)] = [C_{1,k}(2)] + [C_{2,k}(2)] \quad (51)$$

From (45) and (46) it follows that each sub-matrix $[C_{m,k}(2)]$ of size 2×2 could be represented as shown below:

$$[C_{1,k}(2)] = \begin{bmatrix} c_{11}(1,k) & c_{12}(1,k) \\ c_{13}(1,k) & c_{14}(1,k) \end{bmatrix} =$$

$$\text{For } m=1, \quad = \sigma_{1,k} \begin{bmatrix} \cos \theta_{1,k} \cos \theta_{2,k} & \cos \theta_{1,k} \sin \theta_{2,k} \\ \sin \theta_{1,k} \cos \theta_{2,k} & \sin \theta_{1,k} \sin \theta_{2,k} \end{bmatrix}, \quad (52)$$

$$[C_{2,k}(2)] = \begin{bmatrix} c_{11}(2,k) & c_{12}(2,k) \\ c_{13}(2,k) & c_{14}(2,k) \end{bmatrix} =$$

$$\text{For } m=2, \quad = \sigma_{2,k} \begin{bmatrix} \sin \theta_{1,k} \sin \theta_{2,k} & -\sin \theta_{1,k} \cos \theta_{2,k} \\ -\cos \theta_{1,k} \sin \theta_{2,k} & \cos \theta_{1,k} \cos \theta_{2,k} \end{bmatrix}, \quad (53)$$

$$\sigma_{1,k} = \sqrt{\frac{\omega_k + A_k}{2}}, \quad \sigma_{2,k} = \sqrt{\frac{\omega_k - A_k}{2}},$$

$$\theta_{1,k} = \frac{1}{2} \arctg\left(\frac{2\eta_k}{\nu_k}\right), \quad \theta_{2,k} = \frac{1}{2} \arctg\left(\frac{2\delta_k}{\mu_k}\right), \quad (54)$$

$$\omega_k = a_k^2 + b_k^2 + c_k^2 + d_k^2, \quad A_k = \sqrt{\nu_k^2 + 4\eta_k^2}, \quad \nu_k = a_k^2 + c_k^2 - b_k^2 - d_k^2, \quad (55)$$

$$\eta_k = a_k b_k + c_k d_k, \quad \mu_k = a_k^2 + b_k^2 - c_k^2 - d_k^2, \quad \delta_k = a_k c_k + b_k d_k. \quad (56)$$

The sub-matrices $[C_{m,k}(2)]$ of size 2×2 for $k=1,2,3,4$ compose the matrices $[C_m(4)]$, each of size 4×4 (for $m=1,2$):

$$[C_{m,k}(4)] = \begin{bmatrix} [C_{m,1}(2)] & [C_{m,2}(2)] \\ [C_{m,3}(2)] & [C_{m,4}(2)] \end{bmatrix} =$$

$$= \begin{bmatrix} \begin{bmatrix} c_{11}(m,1) & c_{12}(m,1) \\ c_{13}(m,1) & c_{14}(m,1) \end{bmatrix} & \begin{bmatrix} c_{11}(m,2) & c_{12}(m,2) \\ c_{13}(m,2) & c_{14}(m,2) \end{bmatrix} \\ \begin{bmatrix} c_{11}(m,3) & c_{12}(m,3) \\ c_{13}(m,3) & c_{14}(m,3) \end{bmatrix} & \begin{bmatrix} c_{11}(m,4) & c_{12}(m,4) \\ c_{13}(m,4) & c_{14}(m,4) \end{bmatrix} \end{bmatrix}. \quad (57)$$

Hence, the SVD decomposition of the matrix $[X]$ in the first level is represented by two components only:

$$[X(4)] = [C_1(4)] + [C_2(4)] =$$

$$= \begin{bmatrix} ([C_{1,1}(2)] + [C_{2,1}(2)]) & ([C_{1,2}(2)] + [C_{2,2}(2)]) \\ ([C_{1,3}(2)] + [C_{2,3}(2)]) & ([C_{1,4}(2)] + [C_{2,4}(2)]) \end{bmatrix}. \quad (58)$$

In the Second HSVD level ($r=2$), on each matrix $[C_m(4)]$ of size 4×4 is applied 4 times $SVD_{2 \times 2}$. Unlike the preceding 1st level, in the second level the $SVD_{2 \times 2}$ is applied on the sub-matrices $[C_{m,k}(2)]$ of size 2×2 each, whose elements are mutually interlaced and are defined in accordance with the scheme, shown in the right part of Fig. 1. Here the elements of the sub-matrices, on which is applied the $SVD_{2 \times 2}$ in the first and second hierarchical level ($r=1,2$), are tinted in same color. As it could be seen, the elements of the sub-matrices of size 2×2 in the second level are not neighbors, but are placed at one element interval in horizontal and vertical directions. In result of the $SVD_{2 \times 2}$ execution, in the second level each matrix $[C_m(4)]$ is decomposed into two components:

$$[C_m(4)] = [C_{m,1}(4)] + [C_{m,2}(4)] \text{ for } m=1,2. \quad (59)$$

The full decomposition of $[X]$ is then represented as:

$$[X(4)] = [C_{1,1}(4)] + [C_{1,2}(4)] + [C_{2,1}(4)] + [C_{2,2}(4)] =$$

$$= \sum_{m=1}^2 \sum_{s=1}^2 [C_{m,s}(4)], \quad (60)$$

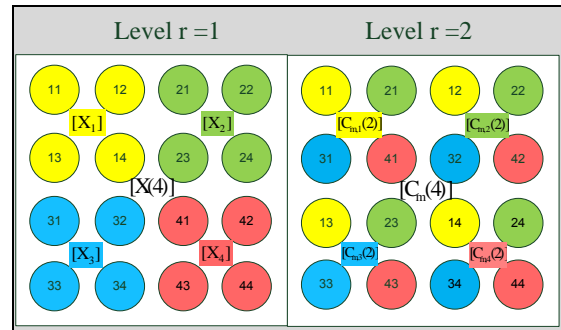


Fig. 1. The elements of the sub-matrices of size 2×2 , over which is applied the $SVD_{2 \times 2}$ in the 1st and 2nd HSVD levels

Hence, the decomposition of the image of size 4×4 comprises four components altogether. When the matrix $[X(8)]$ is of size $2^3 \times 2^3$ ($N=2^3=8$ for $n=3$), three HSVD levels are executed through multiple applying of $SVD_{2 \times 2}$ over the image blocks of size 2×2 . In this case the total number of decomposition components is eight. In the first and second level the $SVD_{2 \times 2}$ is executed in accordance with the scheme, shown on Fig. 1. In the third level, the $SVD_{2 \times 2}$ is applied again on the sub-matrices of size 2×2 . Their elements are defined in a way, similar with this, shown on Fig. 1. The only difference is that the elements of same color (i.e., belonging to same sub-matrix) are moved 3 elements away in horizontal and vertical directions. The presented HSVD algorithm could be generalized for the cases, when the image $[X(2^n)]$ is of size $2^n \times 2^n$ pixels:

$$[X(2^n)] = \sum_{p_1=1}^2 \sum_{p_2=1}^2 \dots \sum_{p_n}^2 [C_{p_1, p_2, \dots, p_n}(2^n)] \quad (61)$$

In this case, the total number of levels is n , and the displacement in horizontal and vertical directions between the elements of the blocks of size 2×2 in the current level r , is correspondingly $(2^{r-1}-1)$ elements for $r=1, 2, \dots, n$.

V. COMPUTATIONAL COMPLEXITY OF THE HIERARCHICAL SVD WITH MATRIX OF SIZE $2^n \times 2^n$

A. Computational complexity of SVD with a matrix of size 2×2

The computational complexity is defined on the basis of (39), taking into account the number of operations multiplication and addition, needed for the preliminary calculation of the components $\omega, \mu, \delta, v, \eta, A, \theta_1, \theta_2, \sigma_1, \sigma_1$, defined by Eqs. (27), (28), (34) and (42). Then:

- The number of multiplications needed for the calculation of (39), is $\Sigma_m = 39$;
- The corresponding number of additions is $\Sigma_s = 15$.

The total number of the needed algebraic operations for the execution of SVD of size 2×2 , is:

$$SS_{SVD}(2 \times 2) = \Sigma_m + \Sigma_s = 54. \quad (62)$$

B. Computational complexity of the hierarchical SVD with a matrix of size $2^n \times 2^n$

The computational complexity of the hierarchical SVD is defined in similar way, as that for the SVD $_{2 \times 2}$. In this case, the number M of the sub-matrices of size 2×2 , contained in the image of size $2^n \times 2^n$, is $2^{n-1} \times 2^{n-1} = 4^{n-1}$, and the number of levels is n .

- The number of SVD $_{2 \times 2}$ in the first level is $M_1 = M = 4^{n-1}$;
- The number of SVD $_{2 \times 2}$ in the second level is $M_2 = 2M = 2 \times 4^{n-1}$;
-
- The number of SVD $_{2 \times 2}$ in level n is $M_n = 2^{n-1}M = 2^{n-1} \times 4^{n-1}$;

The total number of SVD $_{2 \times 2}$ is $M_\Sigma = M(1 + 2 + \dots + 2^{n-1}) = 4^{n-1}(2^n - 1) = 2^{2n-2}(2^n - 1)$ correspondingly, and the total number of algebraic operations for the HSVD of size $2^n \times 2^n$ is:

$$SS_{HSVD}(2^n \times 2^n) = M_\Sigma \cdot SS_{SVD}(2 \times 2) = 55 \times 2^{2n-2}(2^n - 1). \quad (63)$$

C. Computational complexity of SVD with a $2^n \times 2^n$ matrix

For the calculation of matrices $[Y(N)]$ and $[Z(N)]$ each of size $N \times N$, when $N=2^n$, are needed $\Sigma_m = 2^{2n+2}$ multiplications and $\Sigma_s = 2^{n+1}(2^n - 1)$ additions. The total number of operations is:

$$SS_{Y,Z}(N) = 2^{2n+2} + 2^{n+1}(2^n - 1) = 2^{n+1}(3 \times 2^n - 1). \quad (64)$$

In accordance with the analysis, given in [23], the number of the operations $SS(N)$ needed for the iterative calculation of

all N eigenvalues and the eigen N -dimensional vectors of the matrix of size $N \times N$ for $N=2^n$ with L iterations is:

$$SS_{val}(N) = (1/6)(N-1)(8N^2 + 17N + 42) = (1/6)(2^n - 1)(2^{2n+3} + 17 \cdot 2^n + 42), \quad (65)$$

$$SS_{vec}(N) = N[2N(LN + L + 1) - 1] = 2^n[2^{n+1}(2^n L + L + 1) - 1]. \quad (66)$$

From (3) and (4) it follows that two kinds of eigen vectors (\vec{U}_s and \vec{V}_s) have to be calculated, so the number of operations needed for their definition in accordance with (64), should be doubled.

From the analysis of (1) it follows, that:

- The number of multiplications needed to calculate all components is: $\Sigma_m = 2^n(2^{2n} + 2^{2n}) = 2^{3n+1}$;
- The number of additions needed to calculate all components is: $\Sigma_s = 2^n - 1$.

The global number of operations needed for the calculations in accordance with Eq. (1), is:

$$SS_D(N) = 2^{3n+1} + 2^n - 1 = 2^n(2^{2n+1} + 1) - 1 = 2^n(2^{2n+1} + 1) - 1. \quad (67)$$

Hence, the global number of algebraic operations needed to calculate the SVD of size $2^n \times 2^n$, is:

$$SS_{SVD}(2^n \times 2^n) = SS_{Y,Z}(2^n) + SS_{val}(2^n) + 2SS_{vec}(2^n) + SS_D(2^n) = 2^{2n+1}[2L(2^n + 1) + 2^{n-1} + 5] + (1/6)(2^{2n+3} + 17 \cdot 2^n + 42) - 1. \quad (68)$$

D. Determination of the relative computational complexity of the HSVD

The relative computational complexity of the HSVD could be calculated on the basis of (62) and (68), from which is defined the relation below:

$$\psi_1(n, L) = \frac{SS_{SVD}(2^n \times 2^n)}{SS_{HSVD}(2^n \times 2^n)} = \frac{1}{165 \cdot 2^{2n-1}(2^n - 1)} \times \left\{ 3 \cdot 2^{n+1} [2^{n+2}(2^n L + L + 1) + 2^{n+1}(2^n + 3) - 3] + (2^n - 1)(2^{2n+3} + 17 \cdot 2^n + 42) - 6 \right\} \quad (69)$$

The computational complexity of the HSVD is defined by (69). For $n=2, 3, 4, 5$ (i.e., for image blocks, of size 4×4 , 8×8 , 16×16 and 32×32 pixels) the values of $\psi_1(n, L)$ for $L=10$, obtained in accordance with (69), are given in Table 1. For big values of n the relation $\psi_1(n, L)$ does not depend on n and trends towards:

$$\psi_1(n, L)_{n \rightarrow \infty} \Rightarrow (16/165)(3L+1). \quad (70)$$

TABLE 1. THE COEFFICIENT $\psi_1(n, L)$ OF THE RELATIVE LESSENING OF THE COMPUTATIONAL COMPLEXITY OF HSVD TOWARDS THE SVD FOR $L=10$.

n	2	3	4	5
$\psi_1(n, 10)$	5.44	4.14	3.61	3.37

Hence, for big values of n , in the case, when the number of iterations is $L \geq 4$ and $\psi(L) > 1$, the computational complexity of the HSVD is lower than that of the SVD. Practically, the value of L is much larger than four. For the case, given here, $\psi_1(10) = 3$, i.e., the computational complexity of the HSVD is three times lower than that of the SVD.

VI. REPRESENTATION OF THE HSVD ALGORITHM THROUGH TREE STRUCTURE

The presented algorithm for 2-level HSVD with blocks of size 4×4 ($n=2$), represented by (60), could be generalized also for blocks of size $2^n \times 2^n$. In this case, the matrix $[X]$ of each block could be represented by (61).

The number of the HSVD components is n . On the basis of the relations above, on Fig. 2 are shown the corresponding tree structures for the two-level case ($n=2$). As it could be seen from the figures, in accordance with (61) the HSVD algorithm is represented as a binary tree. For a HSVD with a block of size 8×8 , the binary tree should be of levels ($n=3$), while for the tree with three nodes, two levels only are enough. This means, that for the second case the computational complexity is lower.

Each branch of the trees, shown on Fig. 2, has a corresponding eigenvalue $\lambda_{s,k}$, or $\sigma_{s,k} = \sqrt{\lambda_{s,k}}$ for the level 1, and $\lambda_{s,k}(m)$ or $\sigma_{s,k}(m) = \sqrt{\lambda_{s,k}(m)}$ for the level 2 respectively. The total number of branches in the tree from Fig. 2 is equal to 24. A part of the branches in each level could be cut-off, if for them the condition: $\sigma_{s,k} \cup \sigma_{s,k}(m) = 0$, is satisfied, or if their values are close to zero.

To remove one component $[C]$ from given HSVD level, it is necessary all values of σ in this component to be equal or close to zero. In result, the decomposition for the corresponding branch could be stopped before it had reached the last level (n). In this way the HSVD algorithm is adapted in respect to the block contents. In this sense the HSVD algorithm is adaptive and easily adjustable to the requirements of each application.

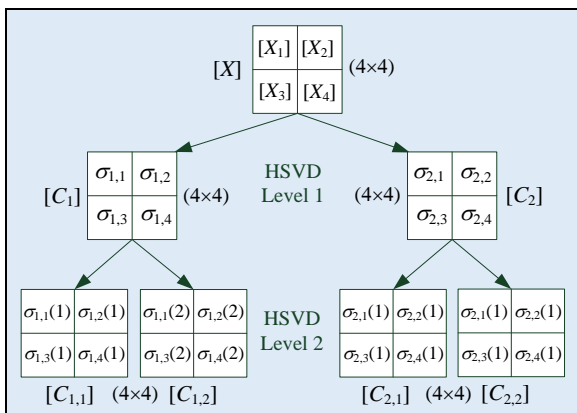


Fig. 2. Representation of the 2-level HSVD algorithm through binary tree

VII. CONCLUSIONS

From the analysis of the presented HSVD algorithm it follows that its basic advantages to SVD are:

1. Its computational complexity, represented as a full tree (without truncation), for a matrix of size $2^n \times 2^n$ ($n=2$) is at least three times lower than that of the SVD, for similar matrix;
2. The HSVD algorithm is represented as a tree structure of n levels, which permits parallel and recursive processing of blocks of size 2×2 in each level. On each block in the corresponding decomposition level is applied the SVD, calculated by using simple algebraic relations;
3. The HSVD algorithm retains the SVD quality to concentrate the basic part of the image energy in the first decomposition components. After removal of the low-energy elements, the restored matrix has minimum mean square error and is an optimal approximation of the original;
4. The tree structure of the HSVD algorithm (a binary tree) facilitates the ability to stop the decomposition in one or more of the tree branches, for which the corresponding eigenvalue is zero, or approximately zero. In result, the HSVD computational complexity is additionally reduced compared to that of the "classic" SVD;
5. The HSVD algorithm could be easily generalized for matrices of any size (not for $2^n \times 2^n$ only). In these cases the matrix should be divided into blocks of size 8×8 , and on each to be applied the HSVD, i.e., will be executed a decomposition of eight components. Beforehand, all incomplete boundary blocks should be expanded through extrapolation. This approach is feasible, when the number of decomposition components, limited up to 8, is sufficient for the application. To increase the number of the HSVD components, the image should be divided into blocks of size 16×16 or larger;
6. The HSVD algorithm opens new opportunities for image processing in various application areas, such as: compression, filtration, segmentation, merging and digital watermarking, extraction of minimum number of features, sufficient for the objects recognition, etc.

REFERENCES

- [1] R. Vaccaro, Ed., SVD and Signal Processing II, Elsevier, New York, 1991.
- [2] M. Moonen and B. de Moor, SVD and Signal Processing III, Elsevier, New York, 1995.
- [3] I. Jolliffe, Principal Component Analysis, 2nd ed., Springer-Verlag, New York, 2002.
- [4] A. Hyvarinen, J. Hurri, P. Hoyer, Natural image statistics, a probabilistic approach to early computational vision, Springer, 2009.
- [5] P. Fieguth, Statistical image processing and multidimensional modeling, Springer, Science+Business Media, 2011.
- [6] E. Carlen, Calculus++, Chapter 3: The Symmetric Eigenvalue Problem, Georgia Tech, 2003.
- [7] H. Andrews and C. Patterson, Singular Value Decompositions and Digital Image Processing, IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-24, 1976, pp. 26–53.
- [8] G. Stewart, The decompositional approach to matrix computation, Computing in Science and Engineering, Vol. 2, No. 1, 2000, pp. 50–59.

- [9] J. Gerbrands, On the relationships between SVD, KLT, and PCA, *Pattern Recognition*, Vol. 14, No. 6, 1981, pp. 375–381.
- [10] D. Kalman, A Singularly Valuable Decomposition: The SVD of a Matrix, *The College Mathematics J.*, Vol. 27, No. 1, Jan. 1996, pp. 2-23.
- [11] S. Orfanidis, SVD, PCA, KLT, CCA, and All That, Rutgers University Electrical & Computer Engineering Dept., *Optimum Signal Processing*, 2007, pp. 1-77.
- [12] R. Sadek, SVD Based Image Processing Applications: State of The Art, Contributions and Research Challenges, *Intern. J. of Advanced Computer Science and Applications*, Vol. 3, No. 7, 2012, pp. 26-34.
- [13] A. Householder, *The Theory of Matrices in Numerical Analysis*, New York: Dover, 1975.
- [14] K. Diamantaras and S. Kung, *Principal Component Neural Networks*, Wiley, New York, 1996.
- [15] A. Levy and M. Lindenbaum, Sequential Karhunen-Loeve Basis Extraction and its Application to Images. *IEEE Trans. on Image Processing*, Vol. 9, No. 8, August 2000, pp. 1371-1374.
- [16] E. Drinea, P. Drineas, P. Huggins, A Randomized Singular Value Decomposition Algorithm for Image Processing Applications, "Proc. of the 8th Panhellenic Conference on Informatics", Y. Manolopoulos and S. Evripidou (Eds.), Nicosia, Cyprus, Nov. 2001, pp. 278-288.
- [17] V. Rokhlin, A. Szlam and M. Tygert, A randomized algorithm for principal component analysis. *SIAM J. Matrix Anal. Appl.* 31, 2009, pp. 1100-1124.
- [18] M. Holmes, A. Gray and C. Isbell, QUIC-SVD: Fast SVD using Cosine trees. *Proc. of NIPS*, 2008, pp. 673-680.
- [19] B. Foster, S. Mahadevan, and R. Wang, A GPU-based Approximate SVD Algorithm. *9th Intern. Conference on Parallel Processing and Applied Mathematics*, 2011, LN in CS, Vol. 7203, 2012, pp 569-578.
- [20] M. Yoshikawa, Y. Gong, R. Ashino, R. Vaillancourt, Case study on SVD multiresolution analysis, CRM-3179, Jan. 2005, pp. 1-18.
- [21] P. Waldemar and T. Ramstad, Hybrid KLT-SVD image compression, *IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing*, IEEE Comput. Soc. Press, Los Alamitos, 1997, pp. 2713-2716.
- [22] M. Aharon, M. Elad and A. Bruckstein, The K-SVD: an algorithm for designing of overcomplete dictionaries for sparse representation, *IEEE Trans. on Signal Processing* 54, 2006, pp. 4311-4322.
- [23] V. Reha, M. Jeyakumar, Singular Value Decomposition based Image Coding for Achieving Additional Compression to JPEG Images, *Intern. J. of Image Processing and Vision Sciences*, Vol. 1, No 2, 2012, pp. 56-61.
- [24] L. De Lathauwer, B. De Moor, and J. Vandewalle, A Multilinear Singular Value Decomposition. *SIAM J. of Matrix Analysis and Applications*, 21, 2000, pp. 1253-1278.
- [25] W. Press, S. Teukolsky and W. Vetterling, *Numerical Recipes in C, The Art of Scientific Computing*, 2nd Ed., Cambridge University Press, 2002.
- [26] K. Dickson, Z. Liu and J. McCanny, QRD and SVD processor design based on an approximate rotations algorithm, *IEEE Workshop on Signal Processing Systems, SIPS*, 2004, pp. 42-47.
- [27] Y. Liu, C. Bouganis, P. Cheung, P. Leong and S. Motley, Hardware Efficient Architectures for Eigenvalue Computation, *Proc. Design, Automation and Test in Europe (DATE'06)*, 2006, pp. 953-958.
- [28] R. Kountchev, K. Nakamatsu, Adaptive multi-level 2D Karhunen-Loeve-based transform for still images, *International Journal of Reasoning-Based Intelligent Systems (IJRIS)*, Vol. 6, Nos. 1/2, 2014, pp. 49-58.
- [29] R. Kountchev, Applications of the Hierarchical Adaptive PCA for Processing of Medical CT Images, *Egyptian Computer Science Journal*, Vol. 37, No 3, May 2013, pp. 1-25.
- [30] R. Kountchev, R. Kountcheva, Adaptive Hierarchical KL-based Transform: Algorithms and Applications, Chapter in "Computer Vision in Advanced Control Systems: Mathematical Theory", M. Favorskaya and L. Jain (Eds.), Springer, Vol. 1, 2015, pp. 91-136.

The Rise of the Robotic Judge in Modern Court Proceedings

H.W.R. (Henriëtte) Nakad-Weststrate LLM

Founder and Director of e-Court
Moersbergselaan 17, 3941 BW Doorn, The Netherlands
E-Mail: henriette.nakad@e-court.nl

H.J. (Jaap) van den Herik

Professor of Computer Science and Law
eLaw, Centre for Law in the Information Society
Leiden University, The Netherlands
E-mail: h.j.vandenherik@law.leidenuniv.nl

A.W. (Ton) Jongbloed

Professor of the Law of Enforcement and Seizure
Utrecht University, The Netherlands
Faculty Law, Economics, Government and Organisation
Molengraaff Institute for Civil Law
E-mail: a.w.jongbloed@uu.nl

Abdel-Badeeh M. Salem

Professor of Computer Science
Head of Artificial Intelligence and Knowledge Engineering
Faculty of Computer and Information sciences
Ain Shams University, Abbasia, Cairo, Egypt
E-mail: absalem@asunet.shams.edu.eg

Abstract—This paper shows an improvement of legal decision-making via digitally produced verdicts. We investigate the use of Artificial Intelligence (AI) in relation to rendering arbitrational verdicts. The data was provided by e-Court, the first private online court of the Netherlands. In our survey the standard debt collection proceedings under Dutch Civil and Procedural law are used as a case study. The introduction of the subject matter is followed by an overview of the key-parameters required by e-Court for rendering a verdict in default cases. The reasoning methodologies of Intelligent Systems in the legal domain are then discussed. Following this discussion we will analyze the nature of the e-Court System to understand how it benefits from the various types of Intelligent Systems. Subsequently, we will discuss the rationale behind the choices made, the legal implications and the handling process within the public courts. Our contribution lies also in the investigation of the characteristics of the e-Court system for rendering default verdicts in debt collection proceedings. In our conclusion we will consider to what extent intelligent systems will be used in the contemporary digital court houses.

Keywords— *intelligent systems, legal decision making, verdicts, rule-based frames, semantic nets, case-based reasoning, e-court, robotic judge, digital judge, arbitration, legislation, e-Court, debt collection, artificial intelligence, legal informatics.*

I. INTRODUCTION

Private court proceedings have played an important role throughout the centuries. As far back as in ancient Egypt there are recordings of arbitration by private courts. The proceedings, which are quite similar to contemporary arbitration, were elaborated by Pharaoh Chephren (26th

century B.C.), also known for the second pyramid of Giza [1 and 2].

In contrast, we deal with *digital private* court proceedings. They are a recent phenomenon. A quarter of a century ago (June 21st, 1991) Jaap van den Herik shocked his audience in his inaugural lecture by addressing the question: “Can computers judge court cases?”. He even wondered whether robotic judges might be better at it than (human) judges ever

would be [3]. A second question was whether computers eventually would deliver such judgments. In one of the final paragraphs, with the title: “2984?”, he stated: “On the basis of these conclusions and beliefs I speculate with you on the future. I will not write science fiction, but rather I want to invoke you to think with me about a future in which the tribunal of reason will be supplemented or supported by tribunals of computers”.

Albeit sooner than expected, his vision became a reality. Within 20 years, on January 11th, 2010 the first online private court in the Netherlands was launched [4 - 10]. It was the first court which offered a fully digitalized court proceedings. Several earlier attempts by the Dutch State, between the late 1980's and 2010, to establish digital public court proceedings had all failed without exception [5-14]. In relation to cases endowed with arguments pro and con, the e-Court verdicts are indeed the result of human reasoning, supported to a large degree by specifically designed software, as forecasted in 1991.

Yet, it did not end there and then. Let us see what happened. Since early 2011, one specific type of verdicts – the e-Court judgments by default in debt collection proceedings – are no longer the product of any human reasoning; the verdicts are rendered as the sole result of AI. Although we may have in mind that the so-called ‘robotic’ or ‘digital’ judge has been in office for a number of years whilst going unnoticed, its appearance in an actual court can be considered a silent revolution in the legal court system.

In our opinion the rise of the robotic judge is a unique development to be distinguished from other developments of our time, such as Crowdsourced Online Dispute Resolution (CODR). To support our opinion we provide a small description. For our definition of the term CODR, we start by using the definition of ODR as provided by Kaufmann-Kohler and Schultz (2004). ODR is “a broad term that encompasses forms of Alternative Dispute Resolution (ADR) and court proceedings which use internet as a part of the dispute resolution process” (Kaufmann-Kohler and Schultz, 2004, p.7). As to the ‘C’ in CODR is refers to ‘the Crowd’. Crowd sourcing has attracted great interest in the academic world, in Europe notably since 2010 and it is even perceived to dominate the future of online dispute resolution [15-26]. Yet, the use of AI functioning on a stand-alone basis, instead of by human reasoning, appears even today a topic of *science fiction*, and in the opinion of many legal professionals a frightening and undesirable future.

In this paper we will set out the relevant key parameters to allow a digital judge to render a verdict by default in arbitration proceedings at e-Court. Then we will focus on the reasoning methodologies for intelligent systems. As a case in point we will investigate a case-study with the following three elements: (i) the Plaintiff is a company, (ii) the Defendant is a consumer, and (iii) the claim amount is a small monetary claim in the domain of debt collection (an unpaid invoice with a maximum of € 1,500).

II. KEY PARAMETERS FOR RENDERING A VERDICT

We consider three different areas: (A) the claim, (B) the costs of debt collection and (C) the course of the proceedings. For each area, there are two classes of key parameters for legal decision making in the sense of rendering a verdict, viz. for (i) the required data, and (ii) the restrictive rules in relation to the use of these data. The relevant required data are to be found for the large part in article 1057 Dutch Code of Civil Proceedings, and article 96, Book 6 Dutch Civil Code. The rules and restrictions follow from the Code of Civil Proceedings, the Civil Code as well as from jurisprudence.

A. Parameters regarding the claim

Regarding the claim, the following data are required for rendering a verdict.

- Claim amount
- Due Date of the claim amount
- Interest over the claim amount
- Interest date
- Full legal names, birth dates and addresses of the Parties

There are at least four restrictive rules in relation to the use of these data.

The first rule is that the contractual basis from which the claim occurred must be clear.

The second rule is that in spite of a Due date of an invoice, the Plaintiff must have sent at least one reminder and a minimum of two collection letters to the Defendant in order for the debt to be payable by a consumer.

The third rule has a relation to the interest. The interest date has to be determined, as well as the percentage of the interest and the proportionality with regard to the claim amount.

The fourth rule is that the calculation of the interest, and other costs (see below, under B) may be affected by the claim amount. Therefore, if the judge does not award the full claim amount as presented by the Plaintiff, the other amounts will be recalculated.

B. Parameters regarding the costs of debt collection

Regarding the costs of debt collection, the following data are required for rendering a verdict.

- Costs of debt collection (made in advance)
- Costs of the writ of summons in which the court proceedings are announced
- Court fee (private court)
- Court fee (public court, to make the binding private verdict enforceable)
- Costs of representation in court

There are at least five restrictive rules in relation to the use of these data.

The first rule is that the cost of debt collection made in advance are limited pursuant to the law.

The second rule is that these costs cannot be claimed, unless the Plaintiff has sent at least one reminder and a minimum of two collection letters to the Defendant in order for the debt to be payable by a consumer.

The third rule is that the cost of the writ of summons are determined by legislation.

The fourth rule relates to the Court fee. The Court fee consists of two elements: (1) the costs of arbitration, which are determined by the private court, and (2) the costs of the public courts to allow execution of the arbitral verdict. Limitations in relation to these costs are found in jurisprudence from the Supreme Court (i.e., verdicts by the Supreme Court). They show that private court proceedings can be considered “unfair” vis-à-vis consumers, if the total costs of the private court exceed the total costs of the public courts for similar cases.

The fifth rule relates to the costs of representation in court. These costs can vary per lawyer. In the Dutch legal system a party can usually only receive a predetermined fixed amount as compensation for the costs. In many legal proceedings this amount is merely a modest contribution in the lawyer’s and court fees.

C. Parameters regarding the course of the proceedings

Regarding the costs of debt collection, the following data are required for rendering a verdict.

- Is the court competent for rendering a verdict in this specific dispute, based on a contract between the parties?
- Was the Defendant duly notified of the oncoming court proceedings by issuing a writ of summons?
- Did the Defendant exercise his right to invoke the competence of the public court for this specific dispute during the four weeks following the writ of summons?
- Were the proceedings held in accordance with the court’s Arbitration Rules?
- Did the Defendant appear in court or was he in default?
- Should the claim nevertheless be rejected because of unlawfulness or unreasonableness?

A number of the six parameters can pose a problem, depending on the factual outcome of the stated question.

For example, if the parties have no contractual clause appointing the private court, the court is not competent and therefore cannot pass a judgment. If the Defendant was not duly notified, the court cannot pass judgment. These parameters are therefore of a “fact finding” nature.

III. REASONING METHODOLOGIES OF ISS

From the knowledge engineering point of view, the main two components in developing an efficient and robust Intelligent System in any domain are (i) the *knowledge base* and (ii) the *inference engines* [27-32].

Ad (i) Concerning the knowledge base there are many knowledge representations and management techniques, e.g. lists, trees, semantic networks, frames, scripts, production rules, cases, and ontologies. The key to the success of such systems is the selection of the appropriate techniques that best fit the domain knowledge and the problem to be solved. The choice depends on the experience of the knowledge engineer.

Ad (ii) Regarding the inference engine, there are many methodologies and approaches of reasoning e.g. automated reasoning, case-based reasoning, commonsense reasoning, fuzzy reasoning, geometric reasoning, non-monotonic reasoning, model-based reasoning, probabilistic reasoning, causal reasoning, qualitative reasoning, spatial reasoning and temporal reasoning. In fact these methodologies receive increasing attention within the AI in law and legal information processing.

Below, we will briefly analyze three distinguished types of Intelligent Systems in the legal domain, previously denoted as Expert Systems. We list (A) Legal Rule-Based Systems, (B) Frames and Semantic Networks, and (C) Case-Based Systems. We will then bring this section to a close under (D) with a discussion of the use of these Intelligent Systems in relation to the presented case-study of default judgment in debt collection proceedings.

A. Survey of Rule-Based Systems

Rule-based systems solve problems by taking an input specification and then “chaining” together the appropriate set of rules from the rule base to arrive at a (new) solution. Given the same exact problem situation, the system will go through exactly the same amount of work and arrive at the new solution. In other words rule-based systems do not inherently learn. In addition, given a problem that is outside the system’s original scope, the system often cannot render any assistance. Moreover, Rule-Based Systems are quite time-consuming to build and maintain. The main reason is that rule extraction from experts is labor-intensive and rules are inherently dependent on other rules, making the addition of new knowledge to the system a complex debugging task [33-35].

Table I shows five examples of Rule-Based Systems for particular legal tasks.

TABLE I. EXAMPLES OF RULE-BASED SYSTEMS FOR PARTICULAR LEGAL TASKS

System	Examples of Rule-Based Systems for particular legal tasks		
	Task	Developing Tools	Rule-Based Systems Site
AUDITOR	Helps a professional auditor evaluate a client's potential for defaulting on a loan	KAS	University of Illinois
DSCAS	Helps contractors analyze the legal aspects of differing site condition (DSC) claims. (Differing Site Condition Analysis System)	ROSIE	University of Colorado
LDS	Assists legal experts in settling product liability cases. (Legal Decision-making System)	ROSIE	The Rand Corporation
SAL	Helps attorneys and claims adjusters evaluate claims related to asbestos exposure. (System for Asbestos Litigation)	ROSIE	The Rand Corporation
TAX-ADVISOR	Assists an attorney with tax estate planning for clients with large estates (greater than \$175,000)	EMYCIN	University of Illinois, and Champaign - Urbana

TABLE II. EXAMPLES OF FRAMES AND SEMANTIC NETS IN LEGAL REASONING AND ARGUMENTATION

System	Examples of Frames and Semantic Nets in legal reasoning and argumentation		
	Task	Developing Tools/K.R. Technique	Site
JUDITH	Helps lawyers reason about civil law cases	FORTRAN/ Relationships	Universities of Heidelberg and Darmstadt
LAS (Legal Analysis System)	Helps lawyers perform simple legal analyses about the interatnional torts of assault & battery	PSL/ Semantic Net	MIT
LRS (Legal Research System)	Helps lawyers retrieve information about court decisions & legislation in the domain of negotiable instrument law, an area of commissioner law that ceals with checks & promissory notes	Knowledge Base/ Semantic Net	University of Michigan
SARA	Helps lawyers analyze decisions governed by discretionary norms	Statistical Tool/ Frames	ROSIE
TAXMAN	Assists in the investigation of legal reasoning and legal argumentation using the domain of corporate tax law	AIMDS/ Frames	EMYCIN

B. Survey of Frames and Semantic Nets

Semantic networks are basically graphical depictions of knowledge that show hierarchical relationships between objects. A semantic network is made up of a number of nodes, which represent objects and descriptive information about those objects. Objects can be any physical items such as a book, car, desk, or even a person. Nodes can also be concepts, events, or actions. The nodes in a semantic network are also interconnected by link or arcs. The arcs show the relationships between the various objects and descriptive factors. Some of the most common arcs are of the is-a or has-a type [36-41].

Table II shows five examples of Frames and Semantic Nets in the domains of legal reasoning and argumentation.

C. Survey of the Case-Based Systems

From a knowledge engineering point of view, a case is a list of features that lead to a particular outcome (e.g., *The information on a legal argument and the associated evidences*). A complex case is a connected set of sub cases that form the problem solving task's structure. Determining the appropriate case features is the main knowledge engineering task in case-based systems. This task involves defining the terminology of the domain and gathering representative cases of problem solving by the expert knowledge engineer. Case-Based reasoning (CBR) is an analogical reasoning method which provides both a methodology for problem solving and a cognitive model of people.

CBR means reasoning from experiences or "old cases" in an effort to solve problems, to give critique on proposed solutions, and explain anomalous situations. It is consistent with observations that psychologist have made in the natural problem solving practice similarly as people do. People tend to be comfortably using the CBR methodology for decision making, in dynamically changing situations and other situations were much is unknown and even when solutions are not clear.

CBR refers to a number of concepts and techniques that can be used to record and index cases and then search them to identify the ones that might be useful in solving new cases when they are presented. In addition, there are techniques that can be used to modify earlier cases to better match new cases and other techniques to synthesize new cases when they are needed [42-45].

From the knowledge engineering point of view, one can summarize the CBR methodology in the following six processes.

1. Assign Indexes: where the features of the new case are assigned as indexes characterizing the event.
2. Retrieve: where the indexes are used to retrieve a similar past case from the case memory (the past case contains the prior solution).
3. Modify: where the old solution is modified to conform to the new situation, resulting in a proposed solution.
4. Test: where the proposed solution is tried out. It either succeeds or fails.
5. Assign and Store: If the solution *succeeds*, then assign indexes and stores a working solution. The successful plan is then incorporated into the case memory.
6. Explain, Repair and Test: If the solution *fails*, then explain the failure, repair the working solution, and test again. The explanation process identifies the source of the problem. The predictive features of the problem are incorporated into the indexing rules knowledge structure to anticipate this problem in the future. The failed plan is repaired to fix the problem, and the revised solution is then tested.

The idea of CBR is becoming popular in developing knowledge-based systems because it automates applications that are based on precedent or that contain incomplete causal models. In a rule-based systems an incomplete mode or an environment which does not take into account all variables could result in either an answer built on incomplete data or simply no answer at all. CBR methodology attempt to get around this shortcoming by inputting and analyzing problem data.

Table III shows seven examples of Case-Based Systems in the legal domain.

III. EXAMPLES OF CASE-BASED SYSTEMS IN THE LEGAL DOMAIN

System	Examples of Case-Based Systems in the legal domain		
	Task	Developing Tools	Rule-Based Systems Site
HYPO	Performs modeling legal argument and adversarial reasoning with cases and hypotheticals in the legal domain	CBR Tool	
LIR	Performs retrieval of legal documents	CBR Tool	
Bank XX	Case-Based legal argument system that retrieves cases and other legal knowledge pertinent to a legal argument through a combination of heuristic search and knowledge-based indexing	CBR Tool	
FLES	Supports the law students in studying the vague concepts in the contracts for the international Sale of Goods. It explains what the meaning of vague legal concept in a query case is	CBR Tool	Tokyo Institute of Technology
LAW-CLERK	Cross-context reminding	CBR Tool	University of Connecticut
GREBE	Exemplar-based Explanation	CBR Tool	University of Texas
JUDGE	Applies the case-based approach to legal reasoning in the context of sentencing convicted criminals	CBR Tool	

D. The nature of the e-Court System for Debt Collection Proceedings

We will now analyze the e-Court System in order to understand the nature of this system and to assess under which type of the Intelligent Systems it can be categorized.

In order to make such an assessment, we have developed a table with an overview of the key tasks in debt collection proceedings under Dutch law. We translate these tasks into system requirements. Finally we analyze what type of Intelligent System is used in the relevant system.

Table IV shows the seven characteristics of the e-Court System for rendering default verdicts in debt collection proceedings.

TABLE IV. CHARACTERISTICS OF THE E-COURT SYSTEM FOR RENDERING DEFAULT VERDICTS IN DEBT COLLECTION PROCEEDINGS

Key Task in Debt Collection Proceedings under Dutch law	Characteristics of the e-Court System for rendering Default Verdicts in Debt Collection Proceedings	
	System Requirement	Nature Rule-Based / Frames and Semantic Nets / Case Based / External ES / Human Intervention
Identify the Parties, and verify their data (birth date, address)	Import the data from the documents (contract, copy of invoices) and verify the data against the state's formal registers.	External ES (court bailiff's) / Human Intervention (court bailiffs)
Establish competence of the court	(a) Review the contract for a forum choice; (b) Establish that the Defendant (i) was duly notified, (ii) did not use his right to evoke competence of the public court	(a) Human Intervention takes place prior to the admission of a Plaintiff to the e-Court system in a principal, pro-active manner rather than in a reactive case-by-case manner; (bi) External ES (court bailiff's expert system, (bii) External ES (court bailiff's expert system
Establish that the proceedings were held in accordance with the court's Arbitration Rules	E-Court System does not allow to deviate from the Arbitration Rules, and the Parties have editing rights for claim/ defense/ reaction/ final defense	Rule Based ES
Select correct template for a Default Verdict	In the absence of an uploaded defense into the e-Court system, the status of the case is Default. The selection of the template is linked to this status.	Rule Based ESs
Award the claimed amounts	Due to the lack of defense the claim is awarded fully	None – Classic calculation tools and models
Produce the digitally signed original, as well as an unsigned copy of the completed verdict	Make the verdict available in PDF, and allow for the original document to be digitally signed.	Rule Based ES
Determine that the claim is not unlawful or unreasonable		Human Intervention takes place prior to the admission of a Plaintiff to the e-Court system in a principal, pro-active manner rather than in a reactive case-by-case manner

Based on the information of Table IV. we may draw three conclusions.

The first conclusion is that the e-Court System makes a limited use of the Rule Based systems, and makes neither use

of the Frames and Semantic Nets, nor of the Case-Based systems.

The second conclusion is that a number of tasks is performed on a pro-active, principle-based approach by human intervention, rather than on a reactive, case-by-case based approach by AI. We will explain the difference by using an example in relation to the establishment by the judge whether or not the court is competent to render a verdict. Let us assume that a health insurance company in the Netherlands wishes to submit its debt collection cases to e-Court for handling it according to the existing legal ruling. The company indicates that there will be approximately 30,000 legal proceedings per annum. As to the first System Requirement, the digital judge will not have to make the assessment on a case by case basis. The assessment is made in an earlier stage, being the moment when e-Court decides whether or not to give the health insurance company access to the e-Court System. There is a plaintiff acceptance policy established, which is similar to the "know your customer" rules and regulations in the financial industry. One topic of investigation is a review of the standard contract used by the Plaintiff, in order to determine whether the standard contract contains a forum choice for e-Court. Following this due diligence of the future Plaintiff, which includes discussions in the field of consumer protection, the Plaintiff will or will not be accepted. This process is performed by human intervention, as e-Court prefers to establish a level of trust and would like be convinced of the integrity and the good faith of the Plaintiff.

The third conclusion is that the success of AI in the legal system will largely depend on finding a well considered path through a minefield consisting of the almost infinite number of technical possibilities, the limited financial resources, and the hindering complexity of the legislation, as well as a legal conservative culture that enhances professional fear and mistrust of applying new ideas in practice.

The introduction of the first digital (i.e., non-human) judge in a legal environment as performed by e-Court has been taken with utmost care. Hence, e-Court started resolving conflicts of a non-complex nature. Here it was soon revealed that even a simple software tool can evoke a huge impact on the legal system. The rationale behind the cautious policy is that a conservative approach of even a small step in technology can show the promise it entails. By doing so, e-Court has successfully averted the danger of falling into the trap of highly complicated, time consuming and expensive development processes that in the end would have resulted in a system far too sophisticated for the tasks ahead. The lessons learned over the past four years allow for further steps in the use of AI in legal decision making.

IV. THE RISE OF THE ROBOTIC JUDGE

In view of the information provided, we acknowledge that there are many benefits of the e-Court robotic judge in relation to our case study. We will restrict ourselves to a discussion of three (evident) benefits of the use of the digital judge, followed

by a presentation of one major legal complication. We will then describe how the public courts processes the verdicts.

A. Three benefits of the e-Court digital judge

The first evident benefit is that the digital judge works fast. In today's world where large numbers of well educated, well organized consumers participate in the economic and legal community as usual consumers by purchasing goods and services, the demand for justice has equally grown to reach a scale that makes the use of ICT-tools a necessity.

The second benefit is that the digital judge can be considered the "most objective judge of the Netherlands", as the judge is impartial and will give rulings without favoring any of the parties involved on the basis of past or present relationships, misplaced empathy, admiration or other subjective influences in the decision making.

The third benefit is that the digital judge works without miscalculations. The software has been designed in such a manner, that all amounts are calculated without the risk of human error.

Here it is recalled that the benefits are based on handling conflicts of a non-complex nature.

B. The major legal complication

There is one major legal complication in relation to the performance of the digital judge as seen from a legal point of view. Despite the benefits of using AI in decision making, Dutch legislation does not provide for the possibility of a digital judge. Its incorporation in the laws and regulations is not to be expected soon, although there is currently some reconsideration. The last fundamental modernization of the arbitration rules has just taken place, and the new arbitration law has come into effect as of January 1st, 2015. The solution to this problem required some legal engineering. The outcome thereof is the situation, whereby the digital judge renders the verdict in the name of the (human) judge. The task of the human judge is therefore limited to a random testing of the verdicts. To date, there has not been one case in which the human e-Court judge was able to improve the verdict by the digital judge.

C. Handling the executional process by the public court

Prior to the execution of an arbitral verdict, one must still obtain a title for execution under Dutch law. These titles are listed in article 430 Dutch Code of Civil Proceedings. In relation to arbitral verdicts the parties will usually seek permission for execution from the public court (article 1062 Dutch Code of Civil Proceedings).

Since 2011, the original, digitally signed verdicts in PDF are sent to the public court, as an attachment (on a CD, USB or other data carrier) to a formal petition (on paper). The court will then print all verdicts on paper, and a court's clerk will manually insert the data, such as the names of the parties, the name of the (human) e-Court judge) in the public court's system. The clerks will then recalculate manually the awarded amounts (claim amount, interest rate, and other costs). This process is manually executed one verdict at a time.

To date, there has not been one case in which the clerks were able to improve the calculations in the verdict. However, there have been examples whereby human error occurred as a result of manual process of the clerks copying all data into the public court's system.

V. CONCLUSIONS

In this paper we examined the use of AI in rendering verdicts by e-Court, the first online private court in the Netherlands. We discussed the legal decision making in the meaning of rendering default verdicts in debt collection proceedings.

We categorized the parameters required for rendering this type of verdicts, followed by a discussion of the Intelligent Systems in legal decision making. After analysis of the nature of the e-Court system, we had to conclude that the e-Court system makes only very limited use of the available Intelligent Systems in legal decision making. We introduced and discussed the rationale behind the cautious approach by e-Court. We then mentioned (1) three benefits of the use of the digital judge, (2) a major legal complication and (3) the manner in which the public courts process these verdicts.

The question arises whether we could argue that the robotic judge has developed from science fiction to a science fact, by coming into existence in this contemporary court house. Our answer to that question would be: "Yes and no".

The answer is "yes", because the verdicts in our case study are indeed generated solely as a result of the – selective, cost efficient and smart - use of AI, without any human reason or intervention involved whatsoever, once the Proceedings have started.

The answer is "no", because the type of cases this robotic judge can handle do not involve the weighing of arguments, the application of case law (jurisprudence) and a decision model in the case of doubt. In other words the robotic judge has not yet come into its full power and existence from a technical point of view.

We will therefore not make the case that the robotic judge has come to its full Artificial Intelligent potential at e-Court. Nevertheless, he has indeed been brought to life, and he has successfully performed its tasks over the past four years. Our overall conclusion would therefore be that we are indeed witnessing the rise of the robotic judge in modern digital private court proceedings.

REFERENCES

- [1] R.A.J. Colenbrander, "Fragments of an investigation into the fundamental being of the office of public notary, especially according the current legislation (Fragmenten van een onderzoek naar het wezen van het notarisambt, in 't bijzonder volgens de tegenwoordige Nederlandsche wetgeving)", Nijkerk: J.J. Malga Jr. 1859, p. 130 e.v.
- [2] A.W. Jongbloed, "Real execution in civil law, Contemplations on real execution under current and future law (Reële executie in het privaatrecht. Beschouwingen over reële executie naar geldend en wordend recht)", (dissertation Nijmegen), Deventer: Kluwer 198, p. 141-142.

- [3] H.J. van den Herik, "Can Computers Judge Court Cases? (*Kunnen Computers Rechtspreken?*)" (inaugural address Leiden 21st June 1991), Arnhem: Gouda Quint 1991.
- [4] H.W.R. Nakad-Weststrate and A.W. Jongbloed, "The digital Highway (De digitale Snelweg)", NL Journal for Litigation (Tijdschrift voor de Procespraktijk) 2010-2, p. 49-51.
- [5] P. E. Ernste, "Court Proceedings at e-Court (Procederen bij e-Court)", NL Journal for Jurisprudence Civil Procedural Law 2010-3, p. 227-232.
- [6] J.C.A. Herstel, "Digital Justice (Digitale Rechtspraak)", Journal for Eastern Holland (Kijk op Oost-Nederland) 2010, p. 1.
- [7] "InternetCourt is launched (Internetrechtbank van start)", the Daily Telegraph (De Telegraaf) 11 januari 2010.
- [8] M. Klompers and M. van Reste, "See you in e-Court?", *Ars Aequi* 2010, p. 301 e.v.
- [9] A.W. Jongbloed, "From both angles: e-Court (Van twee kanten: e-Court)", "NL Judicial Magazine Themis (Rechtsgeleerd Magazijn Themis)" 2014-3, p. 111-116.
- [10] R.C. Hartendorp, "From both angles: e-Court (Van twee kanten: e-Court)", "NL Judicial Magazine Themis (Rechtsgeleerd Magazijn Themis)" 2014-3, p. 117-121.
- [11] M.L. Hendrikse and A.W. Jongbloed, "Civil Procedural law in Practice (Burgerlijk procesrecht praktisch belicht)", Deventer: Kluwer 2005., p. 106.
- [12] A.W. Jongbloed, A.L.H. Ernes, "Rebalancing, Contemplations following the report "Balanced" (Herbalans. Beschouwingen naar aanleiding van het rapport Uitgebalanceerd)", Nijmegen: *Ars Aequi Libri* 2007, p.66.
- [13] "Ministry of Justice loses millions of euros with failed computerization (Justitie verspeelt opnieuw miljoenen euros met mislukte automatisering)" NOS 26 juni 2013, www.nos.nl.
- [14] "Ministry of Justice burns millions of euros by failing ICT-project (Justitie verspilt miljoenen euro's door falend ICT-project)", Elsevier 26 juni 2013, www.elsevier.nl.
- [15] G. Kaufmann-Kohler and T. Schultz (2004). *Online Dispute Resolution: Challenges for Contemporary Justice*, Kluwer LawInternational;
- [16] H.J. van den Herik and D. Dimov, "Towards Crowdsourced Online Dispute Resolution", in: S. Kierkegaard (red.), *Law Across Nations: Governance, Policy & Statutes*, International Association of IT Lawyers (IAITL), 19 september 2011, p. 244-257, available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1933392;
- [17] H.J. van den Herik and D. Dimov, 'Geschilbeslechting door crowdsourcing', *Tijdschrift conflictantering* 2012, p. 19-22; D. Dimov en H.J. van den Herik, 'Използване на краудсорсинг за разрешаване на спорове', *Bulgarian Legal World Magazine* 2012, www.legalworld.bg;
- [18] H.J. van den Herik and D. Dimov, 'Een Crowdsourcing Model voor eBay', in: M. Kreijveld, *Samen Slimmer. Hoe de 'wisdom of the crowds' onze samenleving zal veranderen*, Den Haag: Stichting Toekomstbeeld der Techniek 2012, p. 28-30;
- [19] J. Du Mortier, F. Robben and M. Taeymans, *A Decade of Research at the Crossroads of Law and ICT*, Gent: Larcier 2001;
- [20] A.R. Lodder, 'Conflict resolution in Virtual worlds: General characteristics and the 2009 Dutch convictions on virtual theft', in: K. Cornelius and D Hermann (red.), *Virtual worlds and criminality*, Berlin: Springer 2011, p. 79-93;
- [21] A.R. Lodder and J. Zeleznikow, 'Developing an Online Dispute Resolution Environment: Dialogue Tools and Negotiation Systems in a Three Step Model', *Harvard Negotiation Law Review* 2009-10, p. 287-338;
- [22] A.R. Lodder and J. Zeleznikow, *Enhanced dispute resolution through the use of information technology*. Cambridge: Cambridge University Press 2010. D. Rainey, "Crowdsourced Online Dispute Resolution", Daniel Rainey's blog, 11 September 2009, <http://danielrainey.blogspot.com/2009/09/crowdsourced-onlinedispute-resolution.html>;
- [23] C. Rule and C Nagarajan (2010) *Leveraging the Wisdom of the Crowds: the Ebay Community Court and the Future of online Dispute Resolution*. *ACResolution Volume 2* (Issue 2), 4-7;
- [24] S. Sommers (2006), "On Racial Diversity and Group Decision Making: Identifying Multiple Effects of Racial Composition on Jury Deliberations. *Journal of Personality and Social Psychology* Volume 90 (Issue 4), 597-612;
- [25] Surowiecki, J. (2005). *The Wisdom of the Crowds*. New York City, United States: Anchor Books, a division of Random House;
- [26] N. Tideman (2006). *Collective decisions and voting: the potential for public choice*. Wey Court East, Farnham, Surrey, United Kingdom. Ashgate Publishing Limited;
- [27] Peter Jackson, *Introduction to Expert Systems*, 1998; Cornelius T. Leondes, *Fuzzy Logic and Expert Systems Applications (Neural Network Systems Techniques and Applications)*, 1998; George F Luger, *Artificial Intelligence Structure and Strategies for Complex Problem Saving*, Addison Wesley, 2005
- [28] M. Whitson, Cathy Wu, Pam Taylor, Using an artificial neural system to determine the knowledge based of an expert system, *Proceedings of the 1990 ACM SIGSMALL/PC symposium on Small systems*, p.268-270, March 28-30, 1990, Crystal City, Virginia, USA.
- [29] Joseph C. Gaiarratano, Gary Riley, "Expert Systems", PWS Publishing Co. Boston, MA, USA 1998; Mohammed Almulla , Tadeusz Szuba, Toward a computational model of collective intelligence and its IQ measure, *Proceedings of the 1999 ACM symposium on Applied computing*, p.2-7, February 28-March 02, 1999, San Antonio, Texas, USA.
- [30] Rattapoom Tuchinda, Craig A. Knoblock, Agent wizard: building information agents by answering questions, *Proceedings of the 9th international conference on Intelligent user interfaces*, January 13-16, 2004, Funchal, Madeira, Portugal.
- [31] Cássia T. dos Santos , Fernando S. Osório, An intelligent and adaptive virtual environment and its application in distance learning, *Proceedings of the working conference on Advanced visual interfaces*, May 25-28, 2004, Gallipoli, Italy.
- [32] Andreja Andrić, Vladan Devedić, Marko Andrejić, Translating a knowledge base into HTML, *Knowledge-Based Systems*, v.19 n.1, p.92-101, March, 2006.
- [33] Chakkrit Snae , Pupong Pongcharoen, Automatic rule-based expert system for English to Thai transcription, *Proceedings of the third conference on IASTED International Conference: Advances in Computer Science and Technology*, p.342-347, April 02-04, 2007, Phuket, Thailand.
- [34] Arijit Laha, RAP: a conceptual business intelligence framework, *Proceedings of the 1st Bangalore Annual Compute Conference*, January 18-20, 2008, Bangalore, India.
- [35] Igor Wojnicki, Implementing general purpose applications with the rule-based approach, *Proceedings of the 5th international conference on Rule-based reasoning, programming, and applications*, July 19-21, 2011, Barcelona, Spain.
- [36] P. Hayes (1977) *In Defense of Logic*. In *Proceedings of the fifth International Joint Conference on Artificial Intelligence*, 559-565; P. Hayes (1978) *Naive Physics I: Ontology for Liquids*. In *Formal Theories of the Commonsense World*, eds. J. R. Hobbs and R. C. Moore. Norwood, N.J.
- [37] P. Hayes (1979) *The Logic of Frames*. In *Readings in Knowledge Representation*, eds. R. Brachman and H. Levesque, 288-295; San Mateo, Calif.: Morgan Kaufmann.
- [38] S. Fahlman, D. Touretsky and W. van Roggen, (1981). *Cancellation in a Parallel Semantic Network* In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, 257-263.
- [39] J. Doyle and R. Patil (1989). *Two Dogmas of Knowledge Representation*, Technical Memo, 387B, Laboratory for Computer Science, Massachusetts Institute of Technology.
- [40] W. Hamscher (1991) *Modeling Digital Circuits for Troubleshooting*. *Artificial Intelligence* 51:223-272.

- [41] R. Davis, H. Schrobe, P. Szolovits (1993), What is a Knowledge Representation?, A.I. Magazine, Vol. 14, nr. 1.
- [42] "Case-Based Reasoning Tools from Shells to Object-Oriented Frameworks", Abdrabou, E. A. M. & Salem, A. B. Advanced Studies in Software and Knowledge Engineering- Supplement to the International Journal "INFORMATION TECHNOLOGIES & KNOWLEDGE" Ed. Krassimir Markov, Krassimira Ivanova, Iliia Mitov Sofia: Institute of Information Theories and Applications FOI ITHEA, pp. 37-44,2008.
- [43] "Ideas of Case-Based Reasoning for Keyframe Technique", Hans-Dieter, Abdel- Badea Salem, Bassant Mohamrd El Bagoury, Proceedings of the XVIth International Workshop on the Concurrency Specification and Programming, CS & P 2007,Logow, Warsa, Poland, PP 100-106, 27-29 September 2007.
- [44] "Case Based Reasoning Technology for Medical Diagnosis", Abdel-Badeeh M. Salem, Proceedings of World Academy of Science, Engineering And Technology, CESSE, Venice, Italy, Volume 25, PP 9-13, November 2007.
- [45] M. GR. Voskoglou, A. - B. M. SALEM, "Analogy-Based and Case-Based Reasoning: Two Sides of the Same Coin", International Journal of Applications of Fuzzy Sets and Artificial Intelligence, Vol. 4 , PP 5-51, 2014.

Design of rectangular microstrip antenna with rectangular aperture in the ground plane using artificial neural networks

Siham Benkouda

Electronics Department
University of Frères Mentouri – Constantine 1
Constantine, Algeria
s_benkouda@yahoo.fr

Tarek Fortaki¹, Sami Bedra¹, and Abdelkrim Belhedri²

Electronics Department
¹University of Batna
²University of Frères Mentouri – Constantine 1
Algeria
t_fortaki@yahoo.fr

Abstract— In this paper, we propose a general design of rectangular microstrip antenna with and without rectangular aperture over ground plane, based on artificial neural networks (ANN) in conjunction with spectral domain formulation. In the design procedure, synthesis ANN model is used as feed forward network to determine the resonant frequency and bandwidth. Analysis ANN model is used as the reversed of the problem to calculate the antenna and aperture dimensions for the given resonant frequency, dielectric constant and height of substrate. The spectral domain formulation combined with artificial neural network in the analysis and the design of rectangular antenna to reduce the complexity of the spectral approach and to minimize the CPU time necessary to obtain the numerical results. The results obtained from the neural models are in very good agreement with the experimental results available in the literature.

Keywords—microstrip antenna; artificial neural network; modeling

I. INTRODUCTION

Microstrip antennas (MSAs) are used in a broad range of applications from communication systems to biomedical systems, primarily due to their simplicity, conformability, low manufacturing cost, light weight, low profile, reproducibility, reliability, and ease in fabrication and integration with solid-state devices [1-2]. The main shortcomings of these antennas are narrow bandwidth and low gain. These shortcomings can be overcome by proper design of an antenna, and especially by using proper substrate thickness and dielectric constant as well as a proper way of feeding [3-5].

Several methods [6-9], varying in accuracy and computational effort, have been proposed and used to calculate the resonant characteristics of various microstrip antennas shapes. Generally, there are two methods for analysis of microstrip antenna such as numerical method and analytical method. Despite simple analytical methods giving a good intuitive explanation of antenna radiation properties, exact mathematical formulations involve extensive numerical procedures, resulting in round-off errors and possibly needing final experimental adjustments to the theoretical results [2]. The numerical methods are complicated compared to analytical methods [10]. They are also time consuming and

not easily included in a computer-aided design package [1-2]. On the other hand, commercial software uses computer-intensive numerical methods such as, finite element method (FEM), method of moment (MoM), finite difference time domain (FDTD) method, etc.... But the resulting codes are often too slow for design purposes, since they take a lot of computation time and require large computer resources [11]. To reach to a final optimized structure, it might need several simulations. In order to reduce this time of computation, some commercially available packages are now available with optimizers, but for this also, number of simulations are required [11]. It is well-known that the electromagnetic simulation takes tremendous computational efforts, and the practical measurement is expensive [12].

Currently, computer-aided design (CAD) models based on artificial neural networks (ANNs) have been applied for analysis and synthesis of microstrip antennas in various forms such as rectangular, square, and circular patch antennas [13]. Due to their ability and adaptability to learn, generalizability, smaller information requirement, fast real-time operation, and ease of implementation features [1], neural network models are used extensively for wireless communication engineering, which eliminate the complex and time-consuming mathematical procedure of designing, like the method of

moments [14]. The neural networks in conjunction with spectral domain approach was firstly proposed by Mishra and Patnaik [15], to calculate the complex resonant frequency and the input impedance [16] of rectangular microstrip antenna, this approach is named neurospectral method [8]. This is the main reason for selecting the neurospectral to estimate the resonant frequency and half-power bandwidth of a rectangular microstrip patch over ground plane with rectangular aperture. The analysis model is used to obtain the resonant frequency for a given dielectric material and patch structure, whereas the synthesis model is built to determine patch and aperture dimensions for the required design specifications [12].

The objective of this work is to present an integrated approach based on artificial neural networks and spectral domain approach. We introduce the artificial neural networks in the analysis and synthesis of a rectangular microstrip patch over a ground plane with rectangular aperture to reduce the complexity of the spectral approach and to minimize the CPU time necessary to obtain the numerical results. The neurospectral model is simple, easy to apply, and very useful for antenna engineers to predict both resonant frequency and half-power bandwidth.

II. SPECTRAL DOMAIN FORMULATION

The geometry of the considered structure is shown in Fig.1. We have a rectangular microstrip patch of length L_p along the x direction and width W_p along y direction over ground plane with a rectangular aperture of length L_a and width W_a . Both the center of the patch and the center of aperture have the coordinate value $(x, y) = (0, 0)$. Also, the metallic patch and the ground plane are assumed to be perfect electric conductors of negligible thickness. The dielectric layer of thickness d is characterized by the free-space permeability μ_0 and the permittivity ϵ_0 , ϵ_r (ϵ_0 is the free-space permittivity and the relative permittivity ϵ_r can be complex to account for dielectric loss). The ambient medium is air with constitutive parameters μ_0 and ϵ_0 .

All fields and currents are time harmonic with the $e^{i\omega t}$ time dependence suppressed. The transverse fields inside the substrate region can be obtained via the inverse vector Fourier transforms as [4, 17]

$$\mathbf{E}(\mathbf{r}_s, z) = \begin{bmatrix} E_x(\mathbf{r}_s, z) \\ E_y(\mathbf{r}_s, z) \end{bmatrix} = \frac{1}{4\pi^2} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \bar{\mathbf{F}}(\mathbf{k}_s, \mathbf{r}_s) \cdot \mathbf{e}(\mathbf{k}_s, z) dk_x dk_y \quad (1)$$

$$\mathbf{H}(\mathbf{r}_s, z) = \begin{bmatrix} H_y(\mathbf{r}_s, z) \\ -H_x(\mathbf{r}_s, z) \end{bmatrix} = \frac{1}{4\pi^2} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \bar{\mathbf{F}}(\mathbf{k}_s, \mathbf{r}_s) \cdot \mathbf{h}(\mathbf{k}_s, z) dk_x dk_y \quad (2)$$

where $\bar{\mathbf{F}}(\mathbf{k}_s, \mathbf{r}_s)$ is the kernel of the vector Fourier transforming domain (VFTD) [4, 17]

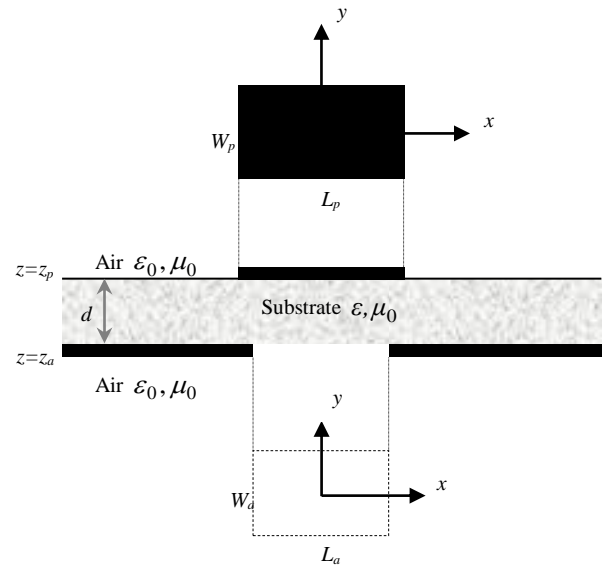


Figure 1. Geometrical structure of a tunable rectangular microstrip patch over a ground plane with rectangular aperture.

$$\bar{\mathbf{F}}(\mathbf{k}_s, \mathbf{r}_s) = \frac{1}{k_s} \cdot \begin{bmatrix} k_x & k_y \\ k_y & -k_x \end{bmatrix} \cdot e^{i\mathbf{k}_s \cdot \mathbf{r}_s}, \quad (3)$$

$$\mathbf{r}_s = \hat{\mathbf{x}}x + \hat{\mathbf{y}}y, \mathbf{k}_s = \hat{\mathbf{x}}k_x + \hat{\mathbf{y}}k_y, k_s = |\mathbf{k}_s|$$

The relation which related the current $\mathbf{j}(\mathbf{k}_s)$, $\mathbf{j}_0(\mathbf{k}_s)$ on the conducting patch (ground plane with rectangular aperture) to the electric field on the corresponding interface $\mathbf{e}(\mathbf{k}_s, z_p)$, and $\mathbf{e}(\mathbf{k}_s, z_a)$ given by

$$\mathbf{e}(\mathbf{k}_s, z_p) = \bar{\mathbf{G}}(\mathbf{k}_s) \cdot \mathbf{j}(\mathbf{k}_s) + \bar{\mathbf{\Psi}}(\mathbf{k}_s) \cdot \mathbf{e}(\mathbf{k}_s, z_a) \quad (4)$$

$$\mathbf{j}_0(\mathbf{k}_s) = -\bar{\mathbf{\Phi}}(\mathbf{k}_s) \cdot \mathbf{j}(\mathbf{k}_s) - \bar{\mathbf{Y}}(\mathbf{k}_s) \cdot \mathbf{e}(\mathbf{k}_s, 0) \quad (5)$$

The four 2×2 diagonal matrices $\bar{\mathbf{G}}(\mathbf{k}_s)$, $\bar{\mathbf{\Psi}}(\mathbf{k}_s)$, $\bar{\mathbf{\Phi}}(\mathbf{k}_s)$, and $\bar{\mathbf{Y}}(\mathbf{k}_s)$ stand for a set of dyadic Green's functions in the vector Fourier transform domain. It is to be noted that $\bar{\mathbf{G}}(\mathbf{k}_s)$ is related to the patch current and $\bar{\mathbf{Y}}(\mathbf{k}_s)$ is related to the aperture field. $\bar{\mathbf{\Psi}}(\mathbf{k}_s)$ and $\bar{\mathbf{\Phi}}(\mathbf{k}_s)$ represent the interactions between the patch current and aperture field. In Equations (4) and (5) the unknowns are $\mathbf{j}(\mathbf{k}_s)$ and $\mathbf{e}(\mathbf{k}_s, z_a)$. Another possible choice in the analysis of microstrip patches over ground planes with apertures is to consider $\mathbf{j}_0(\mathbf{k}_s)$ as unknown instead of $\mathbf{e}(\mathbf{k}_s, z_a)$. It is anticipated, however, that a very large number of terms of basis functions would be

needed for the expansion of the current $\mathbf{j}_0(\mathbf{r}_s)$ on the ground plane with aperture because of the wide conductor area. Hence, it is better to apply the Galerkin procedure to the unknown $\mathbf{E}(\mathbf{r}_s, z_a)$ field at the aperture [17].

The transverse electric field at the plane of the patch and the surface current density on the ground plane with a rectangular aperture can be obtained from Equations (4) and (5), respectively, via the inverse vector Fourier transforms as

$$\mathbf{E}(\mathbf{r}_s, z_p) = \frac{1}{4\pi^2} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \bar{\mathbf{F}}(\mathbf{k}_s, \mathbf{r}_s) \cdot [\bar{\mathbf{G}}(\mathbf{k}_s) \cdot \mathbf{j}(\mathbf{k}_s) + \bar{\Psi}(\mathbf{k}_s) \cdot \mathbf{e}(\mathbf{k}_s, z_a)] dk_x dk_y \quad (6)$$

$$\mathbf{J}_0(\mathbf{r}_s) = -\frac{1}{4\pi^2} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \bar{\mathbf{F}}(\mathbf{k}_s, \mathbf{r}_s) \cdot [\bar{\Phi}(\mathbf{k}_s) \cdot \mathbf{j}(\mathbf{k}_s) + \bar{\mathbf{Y}}(\mathbf{k}_s) \cdot \mathbf{e}(\mathbf{k}_s, z_a)] dk_x dk_y \quad (7)$$

Boundary conditions require that the transverse electric field of Equation (6) vanishes on the perfectly conducting patch and the current of Equation (7) vanishes off the ground plane, to give the following coupled integral equations for the patch current and aperture field:

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \bar{\mathbf{F}}(\mathbf{k}_s, \mathbf{r}_s) \cdot (\bar{\mathbf{G}}(\mathbf{k}_s) \cdot \mathbf{j}(\mathbf{k}_s) + \bar{\Psi}(\mathbf{k}_s) \cdot \mathbf{e}(\mathbf{k}_s, z_a)) dk_x dk_y = \mathbf{0}, \quad \mathbf{r}_s \in \text{patch} \quad (8)$$

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \bar{\mathbf{F}}(\mathbf{k}_s, \mathbf{r}_s) \cdot (\bar{\Phi}(\mathbf{k}_s) \cdot \mathbf{j}(\mathbf{k}_s) + \bar{\mathbf{Y}}(\mathbf{k}_s) \cdot \mathbf{e}(\mathbf{k}_s, z_a)) dk_x dk_y = \mathbf{0}, \quad \mathbf{r}_s \in \text{aperture} \quad (9)$$

The first step in the moment method solution of Equations (8) and (9) is to expand both the patch current $\mathbf{j}(\mathbf{k}_s)$ and aperture field $\mathbf{E}(\mathbf{k}_s, z_a)$ as

$$\mathbf{j}(\mathbf{k}_s) = \sum_{n=1}^N a_n \begin{bmatrix} J_{xn}(\mathbf{r}_s) \\ 0 \end{bmatrix} + \sum_{m=1}^M b_m \begin{bmatrix} 0 \\ J_{ym}(\mathbf{r}_s) \end{bmatrix} \quad (10)$$

$$\mathbf{E}(\mathbf{r}_s, z_a) = \sum_{p=1}^P c_p \begin{bmatrix} E_{xp}(\mathbf{r}_s) \\ 0 \end{bmatrix} + \sum_{q=1}^Q d_q \begin{bmatrix} 0 \\ E_{yq}(\mathbf{r}_s) \end{bmatrix} \quad (11)$$

where J_{xn} , J_{ym} , E_{xp} , and E_{yq} are known basis functions and a_n , b_m , c_p , and d_q are the mode expansion coefficients to be sought. Using the technique known as the moment method [17], with weighting modes chosen identical to the expansion modes, Equations (8) and (9) are reduced to a system of linear equations which can be written compactly in matrix form as

$$\begin{bmatrix} (\bar{\mathbf{U}}^{11})_{N \times N} & (\bar{\mathbf{U}}^{12})_{N \times M} \\ (\bar{\mathbf{U}}^{21})_{M \times N} & (\bar{\mathbf{U}}^{22})_{M \times M} \end{bmatrix} \begin{bmatrix} (\bar{\mathbf{V}}^{11})_{N \times P} & (\bar{\mathbf{V}}^{12})_{N \times Q} \\ (\bar{\mathbf{V}}^{21})_{M \times P} & (\bar{\mathbf{V}}^{22})_{M \times Q} \end{bmatrix} \begin{bmatrix} (\bar{\mathbf{Z}}^{11})_{P \times P} & (\bar{\mathbf{Z}}^{12})_{P \times Q} \\ (\bar{\mathbf{Z}}^{21})_{Q \times P} & (\bar{\mathbf{Z}}^{22})_{Q \times Q} \end{bmatrix} \begin{bmatrix} (\mathbf{a})_{N \times 1} \\ (\mathbf{b})_{M \times 1} \\ (\mathbf{c})_{P \times 1} \\ (\mathbf{d})_{Q \times 1} \end{bmatrix} = \mathbf{0} \quad (12)$$

The elements of the matrix $(\bar{\mathbf{U}})_{(N+M) \times (N+M)}$, $(\bar{\mathbf{V}})_{(N+M) \times (P+Q)}$, $(\bar{\mathbf{W}})_{(P+Q) \times (N+M)}$, and $(\bar{\mathbf{Z}})_{(P+Q) \times (P+Q)}$ are given in [13].

It is easy to show that the entire matrix in Equation (12) is a symmetric matrix. For the existence of a non-trivial solution of Equation (12), we must have

$$\det(\bar{\Omega}(f)) = 0, \quad \bar{\Omega} = \begin{bmatrix} \bar{\mathbf{U}} & \bar{\mathbf{V}} \\ \bar{\mathbf{W}} & \bar{\mathbf{Z}} \end{bmatrix} \quad (13)$$

Equation (13) is the characteristic equation for the complex resonant frequency $f = f_r + if_i$ of the generalized microstrip structure illustrated in Fig.1. f_r is the resonant frequency and $2f_i / f_r$ is the half-power bandwidth of the structure.

In the following section, a basic artificial neural network is described briefly and the application of neural network to the prediction the resonant characteristics of the microstrip antenna are then explained.

III. ARTIFICIAL NEURAL NETWORK

Artificial neural networks (ANNs) have been successfully applied to solve many real world problems, specially the problems which can be hard tracked by expert systems. These networks can predict the relationship between the input and output set without prior knowledge of the process model. The network can solve the problems related with complex engineering systems, difficult electromagnetic computation etc. [18]. In the course of developing an ANN model, the architecture of the neural network and the learning algorithm are the two most important factors. ANNs have many structures and architectures [19-20]. The class of the ANN

and/or the architecture selected for a particular model implementation depends on the problem to be solved.

Multilayer perceptrons (MLP) have been applied successfully to solve some difficult and diverse problems by training them in a supervised manner with a highly popular algorithm known as the error back propagation algorithm [21].

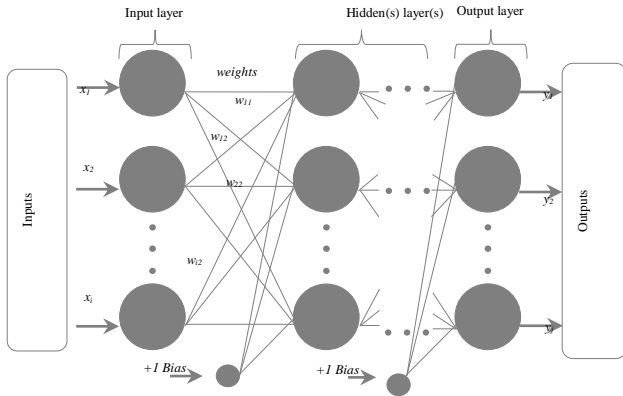


Figure 2. General form of multilayered perceptrons.

As shown in Fig.2, the MLP consists of an input layer, one or more hidden layers, and an output layer. Neurons in the input layer only act as buffers for distributing the input signals x_i to neurons in the hidden layer. Each neuron in the hidden layer sums its input signals x_i after weighting them with the strengths of the respective connections w_{ji} from the input layer and computes its output y_j as a function f of the sum, namely

$$y_j = f\left(\sum w_{ji}x_i\right) \quad (14)$$

Where f can be a simple threshold function or a sigmoid or hyperbolic tangent function [22]. The output of neurons in the output layer is computed similarly. Training of a network is accomplished through adjustment of the weights to give the desired response via the learning algorithms. An appropriate structure may still fail to give a better model unless the structure is trained by a suitable learning algorithm. A learning algorithm gives the change $\Delta w_{ji}(k)$ in the weight of a connection between neurons i and j at time k . The weights are then updated according to the formula

$$w_{ji}(k+1) = w_{ji}(k) + \Delta w_{ji}(k+1) \quad (15)$$

In this work, both Multilayer Perceptron (MLP) networks were used in ANN models. MLP models were trained with almost all network learning algorithms. Hyperbolic tangent sigmoid and linear transfer functions were used in MLP training. The train and test data of the synthesis and analysis ANN were obtained from calculated with spectral model and a computer program using formula given in Section 2. The data are in a matrix form consisting inputs and target values and arranged according to the definitions of the

problems. Using [19-20], two are generated for learning and testing the neural model. The different network input and output parameters are shown in Figure 3 and 4. Some strategies are adopted to reduce time of training and ameliorate the ANN models accuracy, such as preprocessing of inputs and output, randomizing the distribution of the learning data [23], and normalized between 0.1 to 0.9 in MATLAB software before applying training. For an applied input pattern, the arbitrary numbers between 0 and 1 are assigned to initialize the weights and biases [10]. The output of the model is then calculated for that input pattern.

The CPU time taken by the spectral domain to give the both resonant frequency and half-power bandwidth for each input set is more than five minutes; it depends on three initial values used in Muller's algorithm for not seeking of the characteristic equation. All the numerical results presented in this paper we obtained on a Pentium IV computer with a 2.6-GHz processor and a total RAM memory of 2 GB.

In this work, the patch and aperture dimensions of the microstrip antenna are obtained as a function of input variables, which are height of the dielectric material (d), dielectric constants of the substrate (ϵ_r), and the resonant frequency (f_r), using ANN techniques "Fig. 3". Similarly, in the analysis ANN, the resonant frequency of the antenna is obtained as a function of patch (W_p, L_p) and aperture (W_a, L_a) dimensions, height of the dielectric substrate (d), and dielectric constants of the material (ϵ_r) "Fig. 4". Thus, the forward and reverse sides of the problem will be defined for the circular patch geometry in the following subsections.

It should be pointed out that the presence of apertures in the ground plane of microstrip patch antennas unavoidably affects the resonant properties of the antennas. This effect of ground-plane apertures on microstrip patches has been explicitly shown in [24-25,] and [4, 17], where the authors have demonstrated that apertures in the ground plane of rectangular microstrip patches can be used as a way to tune their resonant frequencies [17]. Since ground-plane apertures can play a role in the design of microstrip patch antennas and microstrip patch circuit components. By designer point of view, it is important to give to the calculation of the antenna physical and geometrical parameters the same importance as its resonant characteristics.

Because there is no explicit model that gives the dimension of the patch (ground-plane apertures) directly and accurately and because of the high nonlinearity of the relationship between the resonant frequency and the patch dimension (ground-plane apertures), the reverse modeling is needed [19]. Therefore, this example is very useful for illustrating features and capabilities of synthesis ANN.

A. The forward side of the problem: The synthesis ANN

The input quantities to the ANN black-box in synthesis "Fig. 3" can be ordered as:

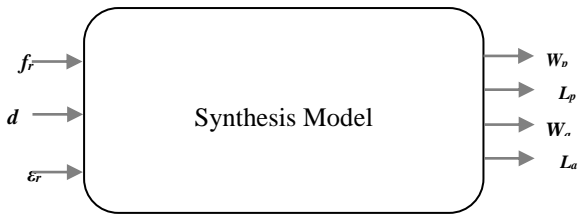


Figure 3. Synthesis Neural model for predicting the patch and aperture dimensions of an antenna with rectangular aperture in the ground plane.

- d : height of the dielectric substrate;
- ϵ_r : effective dielectric substrate;
- f_r : resonant frequency of the antenna.

The following quantities can be obtained from the output of the black-box as functions of the input variables:

- W_p : width of a rectangular patch;
- L_p : length of a rectangular patch.
- W_a : width of a rectangular aperture;
- L_a : length of a rectangular aperture.

B. The reverse side of the problem: The analysis ANN

In the analysis side of the problem, terminology similar to that in the synthesis mechanism is used, but the resonant frequency of the antenna is obtained from the output for a chosen dielectric substrate, patch and aperture dimensions at the input side as shown in “Fig. 4”.

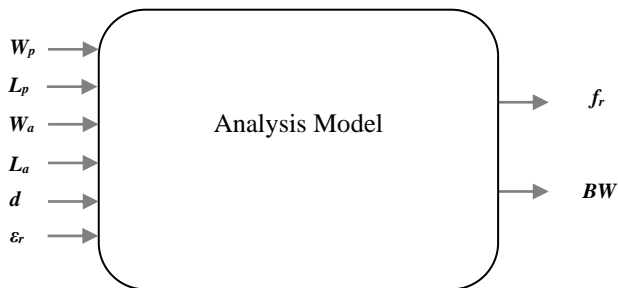


Figure 4. Analysis Neural model for predicting the resonant frequency and bandwidth of rectangular microstrip antenna with rectangular aperture in the ground plane.

To find a proper ANN-based synthesis and analysis models for rectangular microstrip antenna with rectangular aperture in the ground plane, many experiments were carried out in this study. After many trials, it was found that the target of high accuracy was summarized in Table 1.

TABLE 1. COMPARISON OF PERFORMANCE DETAILS OF ANALYSIS AND SYNTHESIS MODEL.

Algorithm details	Neurospectral approach	
	Analysis model	Synthesis model
Activation function	sigmoid	sigmoid
Training function (back-propagation)	trainrp	trainrp
Number of data	250	250
Number of neurons (input layer)	6	3

Number of neurons (2 hidden layers)	12-12	8-10
Number of neurons (output layer)	2	4
Epochs (number of iterations)	5000	10000
TPE (training performance error)	10^{-4}	10^{-4}
Time required	270 min	320 min
LR (learning rate)	0.6	0.5

IV. NUMERICAL RESULTS AND DISCUSSION

In order to determine the most appropriate suggestion given in the literature, we compared our computed values of the resonant frequencies of rectangular patch antennas with the theoretical and experimental results reported by other scientists [26], which are all given in Table 2.

From Table 3 it is observed that the bandwidths of a rectangular microstrip antenna computed by the present approach are closer to the experimental [27], and theoretical values [28-29].

TABLE 2. COMPARISON OF MEASURED AND CALCULATED RESONANT FREQUENCIES OF A RECTANGULAR MICROSTRIP ANTENNA WITH A RECTANGULAR APERTURE IN THE GROUND PLANE; $L_p \times W_p = 34 \text{ mm} \times 30 \text{ mm}$, $\epsilon_r = 2.62$.

Aperture dimension $L_a \times W_a \text{ (mm}^2\text{)}$	Substrate thickness $d \text{ (mm)}$	Resonant frequencies f_r (GHz)	
		Measured [26]	Our results
7×0.7	0.794	2.896	2.901
10×1	3.175	2.750	2.770

In Table 4, the resonant frequencies obtained by the present approach are compared with the previous results [30-31]. The comparison shows that the resonant frequencies computed by the present method are in very good agreement with the measured data for a rectangular patch printed on a single substrate.

TABLE 3. COMPARISON OF THE CALCULATED BANDWIDTH WITH MEASURED AND CALCULATED DATA, FOR A RECTANGULAR MICROSTRIP PATCH ANTENNA WITHOUT APERTURE IN THE GROUND PLANE, $\epsilon_r = 2.33$.

Input parameters (mm)			Bandwidth (%)			
			Measured	Calculated		
W_p	L_p	d	[27]	[28]	[29]	Our results
57	38	3.175	3.12	4.98	3.5	3.75
45.5	30.5	3.175	4.08	6.14	4.0	4.16
17	11	1.524	6.60	8.21	4.8	6.70

TABLE 4. COMPARISON OF CALCULATION AND MEASURED RESONANT FREQUENCIES FOR RECTANGULAR MICROSTRIP ANTENNA WITHOUT APERTURE IN THE GROUND PLANE; WITH $L_p = 25.08 \text{ mm}$, $W_p = 15.438 \text{ mm}$.

The results of the synthesis ANN model and comparison with the targets are given in Table 5. The very good agreement between the values obtained with the model-neuronal synthesis and target values, supports the validity of the neural model. The CPU time taken to calculate the patch dimensions (ground-plane apertures) by using synthesis model is less than a 0.09 second.

In Table 6, we compare our results obtained via the proposed neurospectral model with those obtained using the conventional spectral domain method approach (SDA). As well, to the resonant frequency and half-power bandwidth, we have also shown the CPU time in this table. It is clear that our resonant frequencies and bandwidths coincide with those obtained by the conventional moment method.

Note that, the time required for obtaining the resonant frequency and half-power bandwidth using the neurospectral model is much less in comparison to the spectral domain method.

TABLE .5 RESULTS OF THE SYNTHESIS ANN AND COMPARISON WITH THE TARGETS.

Input parameters			Patch dimension (mm)				Aperture dimension (mm)			
f_r (GHz)	d (mm)	ϵ_r	$W_{-target}$	$L_{p-target}$	W_{p-ANN}	L_{p-ANN}	$W_{-target}$	$L_{a-target}$	W_{a-ANN}	L_{a-ANN}
7.30	0.17	2.22	8.5	12.9	8.485	12.91	1.7	2.6	1.68	2.60
7.98	0.17	2.22	7.9	11.85	7.906	11.84	1.5	2.5	1.48	2.49
3.71	0.79	2.22	20.0	25.0	19.980	24.99	5.0	7.0	5.05	7.06
4.63	1.57	2.33	18.1	19.6	18.121	19.58	3.5	5.5	3.48	5.49
3.96	3.18	2.33	29.5	19.5	19.482	19.49	6.0	4.0	6.03	4.02
7.65	1.52	2.33	17.0	11.0	17.060	11.05	3.4	2.4	3.38	2.39
2.16	1.52	2.50	41.4	41.4	41.385	41.39	6.2	6.2	6.19	6.20
5.07	3.0	2.50	15.3	16.3	15.311	16.28	3.0	3.2	3.03	3.19
6.34	2.42	2.55	11.2	12.0	11.213	12.08	5.6	6.0	5.58	5.98
5.57	2.52	2.55	14.03	14.85	14.052	14.86	2.8	2.8	2.78	2.80
4.42	1.27	10.3	9.1	10.0	9.114	10.06	1.8	2.0	1.81	1.99

TABLE .6 COMPARISON OF OUR RESULTS OBTAINED VIA THE PROPOSED NEUROSPECTRAL MODEL WITH THOSE OBTAINED USING THE CONVENTIONAL SPECTRAL DOMAIN METHOD, WITH $W_p \times L_p = 4 \times 2$ mm².

Input parameters				Conventional method (SDA)			Neurospectral method		
W	L	d	ϵ_r	f_r (GHz)	Bw (%)	CPU Time (min)	f_r (GHz)	Bw (%)	CPU Time (Sec)
(mm)									
2.5	2.5	0.6	2.35	8.956	3.382	5.39	8.972	3.365	0.090
2.5	5	0.8	2.35	8.038	4.221	5.40	8.012	4.178	0.091
5	2.5	1	2.35	8.710	5.621	5.43	8.731	5.642	0.092
2.5	2.5	0.6	3.4	7.531	2.413	5.42	7.562	2.397	0.091
2.5	5	0.8	3.4	6.784	2.931	5.40	6.778	2.894	0.090
5	2.5	1	3.4	7.356	4.063	5.38	7.325	4.023	0.091
2.5	2.5	1.2	10.3	4.365	1.394	5.38	4.352	1.412	0.091
5	2.5	1.4	10.3	4.295	1.645	5.37	4.278	1.637	0.091
2.5	5	1.8	10.3	4.057	2.053	5.41	4.036	1.997	0.090

V. CONCLUSION

In this paper a general procedure is suggested for modeling and design of rectangular microstrip antenna with and without rectangular aperture in the ground plane, using spectral domain approach in conjunction with artificial neural networks. In the design stage, synthesis is defined as the forward side and then analysis as reverse side of the problem. During synthesis of the antenna, it is desirable for the design

Input parameters		Resonant frequency f_r (GHz)			
		Measured		Calculated	
d (mm)	ϵ_r	[30]	[30]	[31]	Our results
0.84	2.2	6.057	6.092	6.15	6.063
1.64	2.2	5.887	5.883	5.89	5.885

engineers to know different performance parameters of an antenna simultaneously, instead of knowing individual parameters, alternatively. Hence, the present approach has been considered more generalized and efficient. The spectral domain technique combined with the ANN method is several hundred times faster than the direct solution. This remarkable time gain makes the designing and training times negligible. Consequently, the neurospectral method presented in this paper is a useful method that can be integrated into a CAD tool, for the analysis, design, and optimization of practical shielded (Monolithic microwave integrated circuit) MMIC devices.

VI. REFERENCES

- [1] K. Guney and N. Sarikaya, "A hybrid method based on combining artificial neural network and fuzzy inference system for simultaneous computation of resonant frequencies of rectangular, circular, and triangular microstrip antennas," IEEE Transactions on Antennas and Propagation, vol. 55, pp. 659-668, 2007.
- [2] A. Kalinli, S. Sagirolu, and F. Sarikoc, "Parallel ant colony optimization algorithm based neural method for determining resonant frequencies of various microstrip antennas," Electromagnetics, vol. 30, pp. 463-481, 2010.
- [3] I. Vilovic, N. Burum, and M. Brailo, "Microstrip antenna design using neural networks optimized by PSO," 21st International Conference in Applied Electromagnetics and Communications (ICECom), 2013, pp. 1-4.
- [4] T. Fortaki, D. Khedrouche, F. Boutout, and A. Benghalia, "Numerical analysis of rectangular microstrip patch over ground plane with rectangular aperture," Communications in numerical methods in engineering, vol. 20, pp. 489-500, 2004.
- [5] M.-H. Ho and C.-I. Hsu, "Circular-waveguide-fed microstrip patch antennas," Electronics Letters, vol. 41, pp. 1202-1203, 2005.
- [6] A. Verma and Nasimuddin, "Multilayer Cavity Model for Microstrip Rectangular and Circular Patch Antenna," Electromagnetics, vol. 24, pp. 193-217, 2004.
- [7] C. Gürel and E. Yazgan, "Resonant frequency of air gap tuned circular microstrip antenna with anisotropic substrate and superstrate layers," Journal of Electromagnetic Waves and Applications, vol. 24, pp. 1731-1740, 2010.
- [8] S. Bedra, S. Benkouda, and T. Fortaki, "Analysis of a circular microstrip Antenna on Isotropic or uniaxially anisotropic substrate Using neurospectral approach," COMPEL: The International Journal for Computation and Mathematics in Electrical and Electronic Engineering, vol. 33, pp. 41-41, 2013.
- [9] D. Guha and Y. M. Antar, Microstrip and printed antennas: new trends, techniques and applications: John Wiley & Sons, 2011.
- [10] T. Khan, A. De, and M. Uddin, "prediction of slot-size and inserted air-gap for improving the performance of rectangular microstrip antennas using artificial neural networks," IEEE Antennas and Wireless Propagation Letters, vol. 12, pp. 1367-1371, 2013.
- [11] S. K. Jain, A. Patnaik, and S. N. Sinha, "Design of custom-made stacked patch antennas: a machine learning approach," International Journal of Machine Learning and Cybernetics, vol. 4, pp. 189-194, 2013.
- [12] W. Zhongbao, F. Shaojun, W. Qiang, and L. Hongmei, "An ANN-Based Synthesis Model for the Single-Feed Circularly-Polarized Square Microstrip Antenna With Truncated Corners," IEEE Transactions on Antennas and Propagation, vol. 60, pp. 5989-5992, 2012.

- [13] M. Aneesh, A. Singh, J. A. Ansari, and S. S. Sayeed, "Investigations for Performance Improvement of X-Shaped RMSA Using Artificial Neural Network by Predicting Slot Size," *Progress in Electromagnetics Research C*, vol. 47, 2014.
- [14] T. Bose and N. Gupta, "Design of an aperture-coupled microstrip antenna using a hybrid neural network," *IET Microwaves, Antennas & Propagation*, vol. 6, pp. 470-474, 2012.
- [15] R. Mishra and A. Patnaik, "Neurospectral computation for complex resonant frequency of microstrip resonators," *IEEE Microwave and Guided wave letters*, vol. 9, pp. 351-353, 1999.
- [16] R. Mishra and A. Patnaik, "Neurospectral computation for input impedance of rectangular microstrip antenna," *Electronics Letters*, vol. 35, pp. 1691-1693, 1999.
- [17] T. Fortaki and A. Benghalia, "Rigorous full-wave analysis of rectangular microstrip patches over ground planes with rectangular apertures in multilayered substrates that contain isotropic and uniaxial anisotropic materials," *Microwave and Optical Technology Letters*, vol. 41, pp. 496-500, 2004.
- [18] P. Samaddar, S. Nandi, S. Nandy, D. Sarkar, and P. Sarkar, "Prediction of resonant frequency of a circular patch frequency selective structure using artificial neural network," *Indian Journal of Physics*, Vol. 88, pp. 397-403, 2014.
- [19] Y. Tighilt, F. Bouttout, and A. Khellaf, "Modeling and design of printed antennas using neural networks," *International Journal of RF and Microwave Computer- Aided Engineering*, vol. 21, pp. 228-233, 2011.
- [20] C. Christodoulou and M. Georgiopoulos, *Applications of neural networks in electromagnetics*: Artech House, Inc., 2000.
- [21] K. Kumar and N. Gunasekaran, "Bandwidth enhancement of a notch square shaped microstrip patch antenna using neural network approach," *International Conference in Emerging Trends in Electrical and Computer Technology (ICETECT)*, Tamil Nadu 2011, pp. 797-799.
- [22] K. Guney and S. Gultekin, "A comparative study of neural networks for input resistance computation of electrically thin and thick rectangular microstrip antennas," *Journal of Communications Technology and Electronics*, vol. 52, pp. 483-492, 2007.
- [23] Z. Raida, "Modeling EM structures in the neural network toolbox of MATLAB," *IEEE Antennas and Propagation Magazine*, vol. 44, pp. 46-67, 2002.
- [24] K. Kawano and H. Tomimuro, "Hybrid-mode analysis of a microstrip-slot resonator," in *IEE Proceedings H (Microwaves, Optics and Antennas)*, Vol. 129, pp. 351-355, 1982.
- [25] K. Kawano, "Hybrid-mode analysis of coupled microstrip-slot resonators," *IEEE Transactions on Microwave Theory and Techniques*, vol. 33, pp. 38-43, 1985.
- [26] M. I. Aksun, S.-L. Chuang, and Y. T. Lo, "On slot-coupled microstrip antennas and their applications to CP operation-theory and experiment," *IEEE Transactions on Antennas and Propagation*, vol. 38, pp. 1224-1230, 1990.
- [27] E. Chang, S. A. Long, and W. F. Richards, "An experimental investigation of electrically thick rectangular microstrip antennas," *IEEE transactions on antennas and propagation*, vol. 34, pp. 767-772, 1986.
- [28] W. C. Chew and Q. Liu, "Resonance frequency of a rectangular microstrip patch," *IEEE Transactions on Antennas and Propagation*, vol. 36, pp. 1045-1056, 1988.
- [29] D. M. Pozar: *PCAAD 3.0. Personal Computer Aided Antenna Design*, Antenna Design Associates, Inc 1996.
- [30] S. Chattopadhyay, M. Biswas, J. Y. Siddiqui, and D. Guha, "Rectangular microstrips with variable air gap and varying aspect ratio: improved formulations and experiments," *Microwave and Optical Technology Letters*, vol. 51, pp. 169-173, 2009.
- [31] HFSS: High Frequency Structure Simulator, Ansoft Corp., 2009.

Design and Implementation of Two Degree of freedom Proportional Integral Derivative Controller

Dr. Raaed Faleh Hassan

Department of Medical Instruments Eng. Techniques

College of Electrical and Electronic Eng. Techniques – Middle Technical University

Baghdad – Iraq

dr-raaed@hotmail.com

Abstract– The paper presents that genetic algorithm can be used for tuning two degree of freedom Proportional Integral Derivative (2DOF – PID) controller. 3rd order plant is considered to be controlled using (2DOF – PID) controller, firstly 3rd order plant has been tested without controller in closed loop system. Secondly the system response is tested with proportional controller then with conventional PID controller, Genetic Algorithm used in both cases in order to obtain optimum response and disturbance rejection. Finally, 2DOF – PID controller is considered for controlling the above plant in closed loop system. The transfer function of this controller is mathematically rearranged in order to have three parameters to be tuned using GA. Simulation results show that 2DOF – PID controller is an effective controller for tracking command signal and disturbance rejection.

Keywords: 2DOF – PID controller, Genetic Algorithm

I. INTRODUCTION

Proportional – Integral – Derivative (PID) controllers are widely used as a core of industrial control applications. This popularity is mainly due to simplicity in determining its parameters and implementation in hardware and software. Despite this popularity, PID controllers cannot provide simultaneously shape the responses to both reference and disturbance signals. This is due to the fact that conventional PID controller has one closed loop transfer function, therefore, it is classified as 1DOF (one degree of freedom) controller [1].

As the PID controller is classified as 1DOF, therefore if the parameters of the PID are tuned to optimize the response of the control system to the command signal, the performance to the disturbance rejection will be deteriorated [1,2].

To solve this problem 2DOF PID controller has been considered in the literature. The dominant feature of the 2DOF PID controller is the two sets of parameters. One of these sets can be used to optimize the performance of the control system to the command signal, while the other sets can be used to optimize the performance against the disturbance [1, 2].

In [1-8], various structures of 2DOF PID controller were proposed and different tuning methods were applied in order to obtain good control performance both in command tracking and disturbance rejection characteristics.

In this paper, Genetic Algorithm GA will be considered for tuning 2DOF PID controller parameters to achieve optimum control performance for a 3rd order plant from command tracking and disturbance rejection points of view.

II. 2DOF PID CONTROLLER

A good performance for both set-point tracking and disturbance rejection can be achieved by using 2DOF PID controller which is also known as ISA – PID controller. The structure contains a standard PID controller in the feedback loop and a pre – filter to the command signal as shown in figure (1) [1].

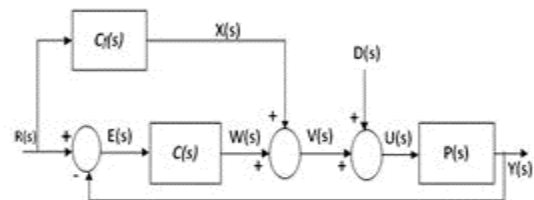


FIG (1): 2DOF PID Control System

Where R(s) is the command signal (reference input), E(s) is the error signal, D(s) is the disturbance and Y(s) is the system response.

The input – output relations of the closed – loop system is given as [1,2]:

$$Y(s) = \frac{P(s)(C(s) + C_f(s))}{1 + P(s)C(s)}R(s) + \frac{P(s)}{1 + P(s)C(s)}D(s) \quad (1)$$

Where:

$$C(s) = k_p \left(1 + \frac{1}{T_i s} + T_d s \right) \quad (2)$$

Which represents conventional PID controller with proportional gain k_p , integral coefficient $\left(\frac{k_p}{T_i}\right)$ and derivative coefficient $k_p T_d$.

and

$$C_f(s) = k_p(\alpha + \beta T_d s) \quad (3)$$

Where α and β are a weighted parameters of feed forward controller, or a pre – filter.

The discrete 2DOF PID controller can be obtained by mapping equations (2) and (3) from s-domain to z-domain using bilinear transformation, therefore:

$$C(z) = C(s) \Big|_{s=\frac{2z-1}{Tz+1}} \quad (4)$$

$$C_f(z) = C_f(s) \Big|_{s=\frac{2z-1}{Tz+1}} \quad (5)$$

Therefore, equation (4) will be

$$C(z) = k_p \left[\frac{\left(1 + \frac{2T_d}{T} + \frac{T}{2T_i}\right)Z^2 + \left(\frac{T}{T_i} - \frac{4T_d}{T}\right)Z + \left(\frac{T}{2T_i} + \frac{2T_d}{T} - 1\right)}{Z^2 - 1} \right] \quad (6)$$

While, equation (5) will be

$$C_f(z) = k_p \left[\frac{\left(\alpha + \frac{2\beta T_d}{T}\right)Z + \left(\alpha - \frac{2\beta T_d}{T}\right)}{Z + 1} \right] \quad (7)$$

Equation (6) can be rearranged to the following form

$$C(z) = \frac{LZ^2 + MZ + (L - 2)}{Z^2 - 1} \quad (8)$$

Where: $L = k_p \left(1 + \frac{2T_d}{T} + \frac{T}{2T_i} \right)$

$$M = k_p \left(\frac{T}{T_i} - \frac{4T_d}{T} \right)$$

and equation (7) can be modified to the following form

$$C_f(z) = \left[\frac{NZ + (N - 2)}{Z + 1} \right] \quad (9)$$

Where: $N = -k_p \left(\alpha + \frac{2\beta T_d}{T} \right)$

The discrete form of the 3rd order plant considered in this paper is:

$$P(z) = \left(\frac{0.0001436z^{-1} + 0.0004951z^{-2} + 0.000104z^{-3}}{1 - 2.464z^{-1} + 2.018z^{-2} - 0.5488z^{-3}} \right) \quad (10)$$

III. DISCRETE – TIME CONTROL SYSTEM

Referring to equations (8 – 10) and fig (1), the set of LTI difference equations are shown below:

$$x(n) = -x(n - 1) + Nr(n) + (N - 2)r(n - 1) \quad (11)$$

$$w(n) = w(n - 2) + Le(n) + Me(n - 1) + (L - 2)e(n - 2) \quad (12)$$

$$v(n) = x(n) + w(n) \quad (13)$$

$$u(n) = v(n) + d(n) \quad (14)$$

$$y(n) = 2.99y(n - 1) + 2.98y(n - 2) - 0.99y(n - 3) + 10^{-7}(1.662x(n - 1) + 6.633x(n - 2) + 1.654x(n - 3)) \quad (15)$$

$$|e(n)| = |r(n) - y(n)| \quad (16)$$

$$IAE = \sum_{k=0}^n |e(k)| \quad (17)$$

Minimization of the IAE represents optimization of the system response to both reference signal $r(n)$ and the output disturbance rejection. To achieve this goal, there are three parameters must be tuned; these are L, M, and N.

IV. GENETIC ALGORITHM

Genetic Algorithm is a stochastic global searching algorithm used to solve complicated problems by simulating the evolutionary course of natural selection and natural inheritance of biology circles. In genetic Algorithm, code space is used to replace problem space, fitness function is regarded as evaluating criterion, code population is regarded as evolution base, selection and genetic mechanism is actualized by genetic operation on individual bit chain of population. A repeated course is formed in this way. The individual of population evolves ceaselessly by recombining some important genes of code bit chain stochastically, and approaches to the optimal gradually till reaching the goal of solving the problem ultimately [9]. In spite of considering GA to be a robust optimization algorithm, it has some drawbacks. The main drawback is the GA cannot assure constant optimization response time, and it is unreasonable to use GA for on - line controls in real system

V. OPTIMIZATION PROCESS

In this section the methodology for tuning the optimization of 2DOF – PID controller is described.

Firstly, the plant in eq.(10) has been considered in closed – loop control system with proportional controller only. GA used to obtain the optimum gain for the proportional controller, in this situation the optimization process performed for tuning one parameter.

1DOF – PID controller used for controlling the plant in closed loop control system, the optimization process in this case will has three parameters to be tuned.

Finally, the set of difference equations (11 – 17) has been implemented as a Matlab function file in order to calculate the fitness function which is Integral Absolute Error (IAE). The task of GA algorithm is to minimize this fitness function by selecting the three unknowns L, M and N.

VI. SIMULATION RESULTS

Fig (2) shows the response of the closed loop system of the 3rd order plant without controller. From his figure it can be seen that the system response isn't track the command signal and there is significant error.

Genetic Algorithm has been used to determine the optimum gain needed to minimize the error between the command signal and the system response.

Fig (3) shows that only adding gain to the system isn't sufficient because this gain produces high peak overshoot (more than 20%) and a hard oscillation.

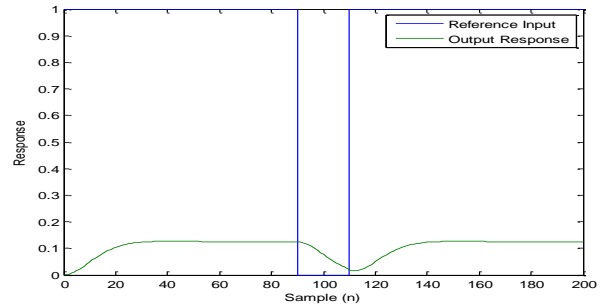


FIG (2): 3rd order plant without controller

The results of optimization process using GA are summarized in table 1

Table 1

No. of Generation	No. of evaluation	Best fitness
292	586000	41.4932

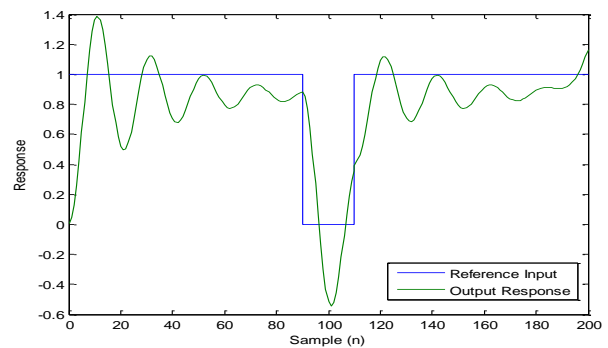


FIG (3): Closed Loop control system with proportional controller.

According to the previous results and in order to overcome the challenges appear, 1 DOF – PID controller has been considered.

The task of GA is to tune three parameters k_p , k_i & k_d of the PID controller for obtaining the optimum response to the command signal.

Fig (4) shows system response to the command signal. The result indicates that the response has very high peak overshoot and steady state error.

The results of optimization process are summarized in table 2.

Table 2

No. of Generation	No. of evaluation	Best fitness
1000	2002000	49.2411

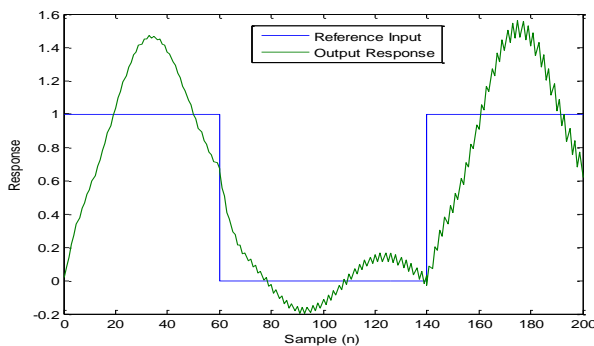


FIG (4): Closed Loop control system with 1DOF - PID controller.

Now, conventional PID controller has been tested under the presence of output disturbance. Fig (5) shows the system response for unit step input and finite duration of output disturbance. From these results, it can be noticed that the control system has acceptable disturbance rejection but tracking of the command signal still has high peak overshoot.

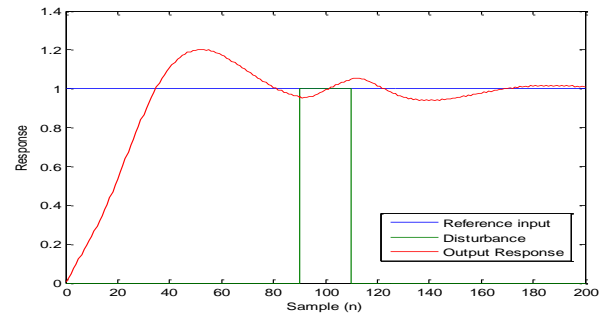


FIG (5): Disturbance rejection of closed loop control system with 1DOF PID Controller

The optimization process results are summarized in table 3.

Table 3

No. of Generation	No. of evaluation	Best fitness
1000	2002000	49.2411

2DOF PID controller is considered in order to improve system response for tracking the command signal and disturbance rejection. Fig (6) shows simulation results of the system response to the command signal.

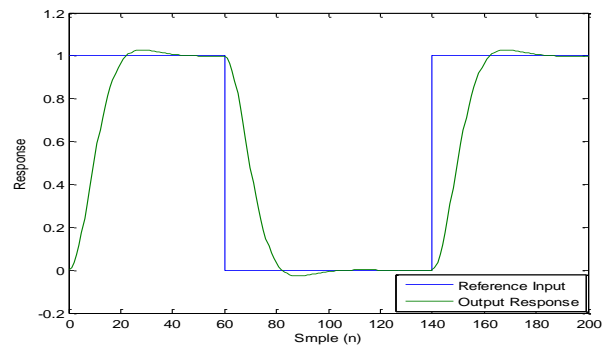


FIG (6): Closed Loop control system with 2DOF - PID controller.

The optimization process results are summarized in table 4.

Table 4

No. of Generation	No. of evaluation	Best fitness
235	472000	32.2484

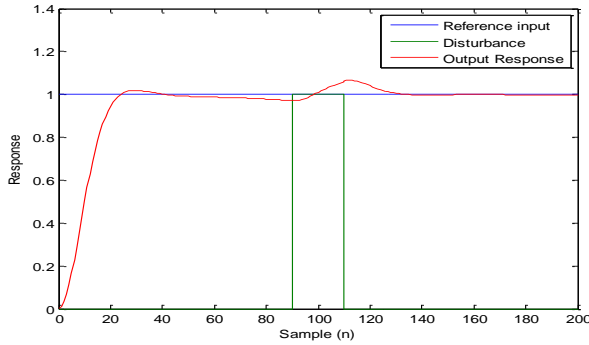


FIG (7): Disturbance rejection of closed loop control system with 2DOF PID Controller

2DOF PID controller is also considered for controlling the plant to achieve both command signal tracking and output disturbance rejection.

Fig (7) shows simulation results of this situation, it is clear from these results that the using of 2DOF PID controller improved the performance of the control system for tracking command signal and output disturbance rejection.

The results of optimization process are summarized in table 5.

Table 5

No. of Generation	No. of evaluation	Best fitness
291	584000	12.4443

Table 6 shows the performance summary of different controllers.

Table 6

Type of Controller	Fitness function (IAE)	
	Command Signal	Disturbance Rejection
Proportional K_p	41.4932	-----
1DOF PID	49.2411	27.8699
2DOF PID	32.2484	12.4443

VII. CONCLUSIONS

In this paper genetic algorithm has been used as an optimization tools for tuning different controller structures for controlling 3rd order plant. Proportional, conventional PID and 2DOF – PID controllers are considered for controlling the above plant. The parameters of these controllers are tuned using genetic algorithm and from simulation results it can be seen that only 2DOF – PID controller able to make the system to keep track of the command signal as well as reject the output disturbance.

REFERENCES

- [1] Mitsuhiro Araki and Hidefumi Taguchi “Two – Degree – of – Freedom PID controllers”, International Journal of Control, Automation and Systems, Vol. 1 No. 4, December 2003.
- [2] Jing – Gang Zhang, Zhi – yan Liu and Run Pei “ Two – Degree – of – Freedom PID control with Fuzzy Logic Compensation”, proceeding of the first International conference on machine learning and Cybernetics, Beijing, 4 – 5 November 2002.
- [3] Takao Sato, Akira Inoue and Toru Yamamoto “ Two – Degree – of – Freedom PID controller based on Extended Generalized Minimum Variance Control” International Journal of Innovative computing Information and Control, Vol. 4, No. 12, December 2008.
- [4] Gish Herjolfsson, Anna Soffia Hauksdottir and Sven P. Sigurdsson “ A Two – Stage optimization PID Algorithm” IFAC conference in PID control, Brescia (Italy), March 28 – 30, 2012.
- [5] Eiji Takegami, Kohji Higuchi, Kazushi Nakano, et. al.” The Method for Determining Parameters of Approximate 2DOF Digital Controller for Robust Control of DC – DC Converter” ECTI Transactions on Electrical Eng., Electronics and communications, Vol.4. No.1, February 2006.
- [6] Tohru Kawabe “ Robust 2DOF PID controller Design of Time – Delay Systems Based on Evolutionary computation” 4th WSEAS International conference on Electronics, Control and signal Processing, Miami, Florida, USA, 17 – 19 November 2005 (pp 144 – 149).
- [7] Humberto M. Mazzini, Fabio G. Dos Santos “ Two – Degree – of – Freedom PID Control for Integrating process” XVIII congress, Brazil, 12 – 16 September 2010.

- [8] Takao Sato, Akira Inoue and Yoichi Hirashima, “Self-Tuning Two-Degree-of-Freedom PID Controller Reducing the Effect of Disturbances” Proceedings of the American Control Conference Anchorage, AK May 8-10, 2002.
- [9] Tai-Shan Yan, Yong-Qing Tao, and Du-Wu Cui, “Research on Handwritten Numeral Recognition Method Based on Improved Genetic Algorithm and Neural Network”, Proceedings of the 2007 International Conference on Wavelet Analysis and Pattern Recognition, Beijing, China, 2-4 Nov. 2007.

Experiments with Simulated Humanoid Robots

Hans-Dieter Burkhard
Humboldt University Berlin
Institute of Informatics
Berlin, Germany
nao-team@informatik.hu-berlin.de

Abstract— Experimenting with real robots is limited by the available resources: Complex hardware is costly, and it needs time and experience for setup and maintenance. Simulated robots can be used as alternative. Our RoboNewbie project is a basic framework for experimenting with simulated robots. It serves as an inspiration for beginners, and it provides room for many challenging experiments. The RoboNewbie agents run in the simulation environment of SimSpark RCSS, the official RoboCup 3D simulator, where the simulated robots are models of the humanoid Robot NAO of the French Company Aldebaran. Different example agents provide easily understandable interfaces to simulated sensors and effectors of the robot as well as simple control structures. The framework has been successfully used at different courses where the participants needed only few hours to understand the usage of the framework and to develop own agents for different tasks.

Keywords— Robotics, e-Learning, Simulated Robots, RoboCup

I. INTRODUCTION

Understanding grows with active commitment: to "do" something, to master it, provides a deeper understanding. Experiencing with own experiments is an important prerequisite for studies in Robotics and Artificial Intelligence. But experimenting with real robots is difficult not only because of expensive hardware. Maintaining the robots and set ups for experiments are very time consuming even for experienced people. Experiments at home as needed for e-learning would require a deep technical understanding by the students, i.e. experiences that they are just going to learn. So it is not surprising that simple hardware is still broadly used in robot experiments, hardware which is far behind the recent technical developments, not to talk about e.g. complex humanoid robots. The collection of papers [1] can be understood as an illustration of that statement.

Simulated robots in simulated environments are widely used as an alternative for complex hardware. They are often simulations of existing robots and serve for preliminary programming, tests and evaluations. Because of the "reality gap", the transfer of programs from simulated to real robots is a non-trivial task [2]. Reducing the reality gap needs increasing efforts in the simulation and leads again to complex systems which need more efforts by the user. The trade-off must be handled carefully for systems better suited for beginners.

The RoboCup community has more than 15 years of experiences with real and simulated robots in the field of soccer playing robots [3]. Soccer playing robots have been established as a challenging test field for the progress in scientific research and technical developments. Robots have to be able to control their bodies and their motions according to soccer play, they must perceive a dynamically changing

environment and they have to choose successful actions out of many options in real time. They have to cooperate with team mates and to pay attention to opponents. Several thousand scientists and students are participating in the annual RoboCup competitions in different leagues with different types of real and simulated robots. The humanoid robot Nao of the French Company Aldebaran [4] is used in the Standard Platform League, while its simulated version is used in the 3D-Simulation League. The official SimSpark RoboCup 3D Soccer Simulation (SimSpark RCSS) [5] provides an excellent environment for experiments with simulated complex robots (see Section III). It provides a physical simulation using ODE [6] for the body dynamics of the robot Nao and the soccer environment.

Our RoboNewbie Project is a basic framework based on JAVA for the development of simulated humanoid robots. It provides easily understandable interfaces to simulated sensors and effectors of the robot as well as a simple control structure. It runs in the environment of the SimSpark RCSS, thus it can but need not be used for soccer playing robots. Users can develop their own motions, e.g. for dancing, gymnastics or kicking a ball.

The RoboNewbie Project implements some kind of "minimalistic approach" with respect to Robotics. Users are able to start without special knowledge about robots. They can learn by their own experiences about the basic concepts of perception, motion, control, synchronization, and integration. All related program code in RoboNewbie is understandable from simple principles without further knowledge. That concerns the structure of the code as well as the underlying computational methods. As soon as users learn more about Robotics, they will be able to extend the programs accordingly, e.g. concerning complex motions or world modeling.

Moreover, the framework has also good potentials for the research on foundations. e.g., on computational models as well as on different problems in cognitive science. It can be useful in verifying models and in gathering large data sets for experiments in data mining.

The paper is organized as follows: After an overview about the concept and the downloadable resources of the RoboNewbie project, it gives a short overview about SimSpark RCSS. The communication between the RoboNewbie agents and SimSpark RCSS is described next. The main part of the paper in Section V discusses the details of the RoboNewbie framework, and the paper ends with results of practical evaluations and conclusions.

II. THE ROBONEWBIE PROJECT AND ITS RESOURCES

The main goal of the RoboNewbie Project is to provide an uncomplicated starting point to the programming of complex robots with minimal requirements and pre-knowledge. The users are only supposed to have some programming background (Java) and some technical/mathematical understanding. More knowledge about Robotics can be provided in parallel to the exercises with RoboNewbie, e.g. in introductory tutorials (as was already done) or by e-Learning material.

The objective behind RoboNewbie is the realization of the following requirements:

- Holistic view on robots: For beginners, it is more appealing to see a robot behave like a human than to test and calibrate the behavior of a sensor. Of course, when dealing with more complex tasks, users will experience the need to have better knowledge about the usage of sensors and actuators, and then they may draw their own conclusions.
- Motivating scenario: Application fields from daily life with known properties and rules are well suited. Robots which imitate human skills are especially motivating.
- Scalable tasks: Inexperienced users should have no difficulties to perform first steps with own experiments and later move to more complex tasks with unlimited challenges.
- Low requirements: The usability would be restricted if people need pre-knowledge on Robotics or if they are supposed to have deep knowledge in hardware and software. Basic programming skills and interests in mathematics and natural sciences should be sufficient.
- Low costs: The costs of a learning system include money and efforts for purchase, set up, and maintenance, respectively. They should be as low as possible to permit a broad usage.

The users of the RoboNewbie project can find all materials on the web page of Berlin United -- Nao Team Humboldt [7].

Besides links to RoboCup, Nao (Aldebaran) and the SimSpark-Wiki, it contains resources for download:

- Description of installation and first steps.
- Sources of RoboNewbie agents programmed in JAVA 7 and prepared for usage under Netbeans.
- “Quick start tutorial”: Introduction to the features and the usage of the agent.
- Motion Editor for the design of Keyframe Motions (needs JAVA 3D to be installed).
- SimSpark RoboCup 3D Soccer Simulation (SimSpark RCSS) for Windows with an introduction to SimSpark RCSS as far as needed for RoboNewbie.

All provided code is open source. Some parts of the RoboNewbie code use code of the RoboCup team magmaOffenburg [8].

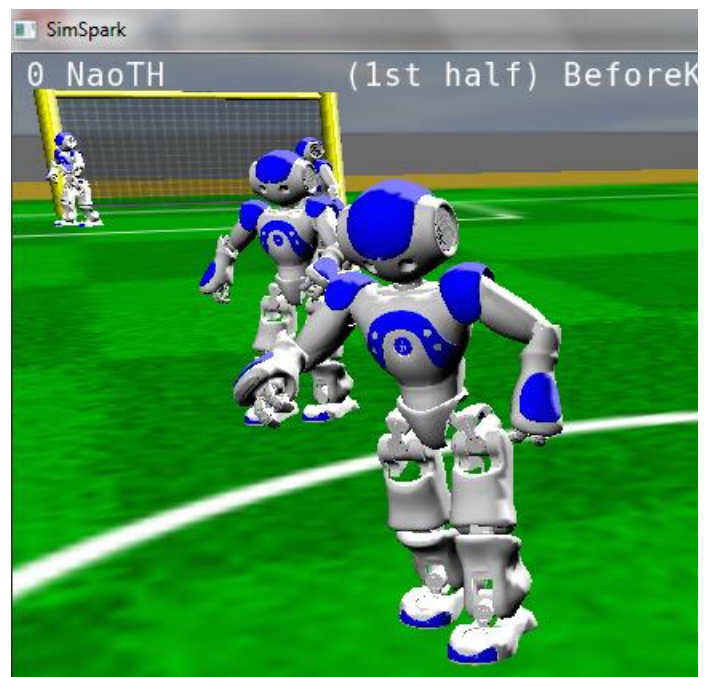


Figure 1: SimSpark RoboCup 3D Soccer Simulation as used in the RoboCup competitions. The field size is reduced for RoboNewbie.

III. SIMSPARK ROBOCUP 3D SOCCER SIMULATION

SimSpark RCSS is developed and used by the RoboCup community in the 3D simulation league. SimSpark is a generic physical multi agent simulator system for agents in three-dimensional environments. It uses the Open Dynamics Engine (ODE [6]) for detecting collisions and for simulating rigid body dynamics. ODE allows accurate simulation of the physical properties of objects such as velocity, inertia and friction.

The Simulator SimSpark RCSS consists of two programs (server for simulation and monitor for visualization and interaction) together with configuration files. It models a soccer field with the player bodies (adapted from the robot hardware of Nao) and the ball. It also controls the rules of the soccer game, i.e. it controls the game according to the decisions of a simulated referee.

SimSpark RCSS can be used as open source software. This was also an important criteria for its usage. It can be downloaded from [5] for different platforms. A complete preconfigured version for Windows 7 is provided for RoboNewbie which can be downloaded from the RoboNewbie web page [7]. Nevertheless, the RoboNewbie agents run with SimSpark RCSS under other platforms, too.

By some small changes in the configuration files, the soccer rules are simplified for first usages with RoboNewbie. The SimSpark RCSS project itself is constantly evolving according to the progress in the RoboCup initiative. The version (compiled in June 2012) on the RoboNewbie web pages serves for stable usage and avoids potential incompatibility problems by new RoboCup versions.

SimSpark RCSS is documented in a Wiki [5] with download links to the latest versions as used in the competitions. The Wiki documentation is thought to represent the actual state of the simulator by continuous updates. But since different developers are volunteering in parallel on different tasks in the project, the structure of the Wiki is not always optimal, and occasionally some outdated information is still present. Moreover, the Wiki is directed to experienced users. This makes it sometimes difficult to understand for novices. According to our experiences (cf. Section VI), the deeper knowledge is not needed by beginners.

To provide an easy access, the downloads of the RoboNewbie project contain an introduction to SimSpark RCSS which refers to the provided version (as described above). It gives the user an overview about

- Simulation using SimSpark RCSS: The SoccerServer and the Monitor.
- The Nao-Model used by SimSpark RCSS.
- Communication between Agents and SimSpark RCSS (with explanations of the message formats as background information).
- \Synchronization between SimSpark RCSS and the Agents.
- Monitor and User Interface.
- Running a Game.

Actually, our description of SimSpark RCSS provides also some "background" information which is not needed for beginners, e.g. details about the message formats. Since RoboNewbie permits an easy and direct access to the items of messages like sensor values and motor commands, the syntax of messages must not be known by users. Nevertheless, we

have included the information for deeper understanding of RoboNewbie in case of interest.

IV. COMMUNICATION BETWEEN AGENTS AND SIMSPARK

SimSpark RCSS implements the soccer environment including the bodies of the Nao robots. It models all physical interactions between players, ball and environment. The agents implement the control of the players.

The interface between the physical environment and the control of real robots is constituted by sensors and actuators: Robots perceive the world by sensory data (e.g. by vision, accelerometer, force sensors etc.), and influence the world by their actuators (motors, voice etc.).

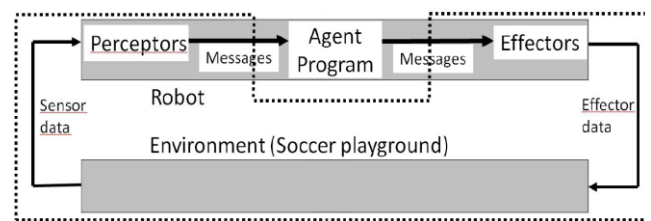


Figure 2: Simulation scheme with message transfer between the agent and the simulated physical world. The simulated physical world consists of the robot hardware and the environment. The agent controls the actions of the robot. Messages contain data of the preceptors (sensors) and effectors (actuators).

In simulation (Figure 2), the sensory data are calculated by the simulator according to the situation in the simulated world (e.g. observable objects) and sent via messages to the agent. Then, like a real robot, the agent can update its belief about the situation and decide for actions it wants to perform. A real robot would then activate its actuators (e.g. motors at the joints) to perform the intended actions. In simulation, the agent communicates with SimSpark RCSS again by messages which transmit the effector commands. Both are synchronized by a communication cycle of 20 milliseconds.

In SimSpark RCSS, the message transfer is optimized for minimizing the server and the traffic load: All sensory data are packed in one server message to be sent at the beginning of a communication cycle. Vice versa, the agent can send all action commands by a single agent message before the end of a cycle. This trick makes it possible to run several agents together with the simulator even on a simple laptop.

The message formats follow a special syntactic scheme based on symbolic expressions (S-expressions). As a consequence of collecting data into one message, the preparation of the data in an agent needs more efforts than in a real robot. It is a special feature of the RoboNewbie agents that this preparation is hidden from the user: The agents provide special getter- and setter-methods which allow the access to the perceptor (sensor) data and the setting of effector (actuator) commands in a similar way as in a real robot.

The interaction between the server and the agents works as follows (see Figure 3):

1. At the beginning of a cycle at a time t , the server sends individual server messages with sensations to the agents.
2. During this cycle, the agents can decide for new actions depending on their beliefs about the situation.
3. Before the end of this cycle, the agents should send their agent messages to the server for desired actions.
4. The server collects the messages of all agents and calculates the resulting new situation (poses and locations of the players, ball movement etc.) according to the laws of physics and the rules of the game. This is done during the following cycle at time $t+1$, i.e., the server message sent at the beginning of this cycle regards the situation calculated in the previous cycle at time t . We have a reaction delay as in reality (see Figure 3).
5. At the beginning of the subsequent cycle, at time $t+2$, the sensor data in the server message is based on the effects of the actions at time $t+1$ which were chosen by the agent according the information from time t .

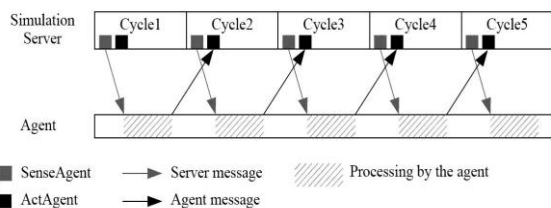


Figure 3: Message exchange during the server cycles.

A special feature of SimSpark RCSS is the use of so-called perceptrors instead of sensors. The perceptror data can be regarded as already pre-processed sensor data. For example, the image data from the camera are not presented by a pixel matrix. Instead, the vision perceptror sends a collection of observable objects with egocentric coordinates relatively to the camera of the observing agent. In a similar way, action commands of the agent are encoded as so-called effector values and sent to the server which translates them to motor control commands. The calculation of perceptror values and the interpretation of effector values are part of the simulator, too.

On the agent side, a server message has to be parsed for the contained perceptror values, and the action commands have to be collected to the agent message. Both constitute a significant burden for a beginner while it provides only few insights to robotics. The RoboNewbie users need not to care about that, because the RoboNewbie agents do all this work in the background.

Besides some effectors related to initial connection with SimSpark RCSS, there are hinge joint effectors for each of the

22 hinge joints (see Figure 4) and a say perceptror (as of a loudspeaker with limited capacity). The following perceptrors are available in SimSpark RCSS (for details see the Wiki or our SimSpark description):

- Vision Perceptror (as of a camera in the center of the head).
- Hinge Joint Perceptrors at each of the 22 hinge joints.
- Accelerometer in the centre of the torso.
- GyroRate Perceptror in the centre of the torso.
- Force Resistance Perceptror at each foot.
- Hear Perceptror (as of a directed microphone with limited capacity).
- Game State Perceptror (reporting the actual game state of the soccer match).

V. ROBO NEWBIE FRAMEWORK

The RoboNewbie framework offers a comfortable interface for agents interacting with SimSpark RCSS. It includes sample agents which illustrate basic concepts and methods of Robotics and Artificial Intelligence. Users can start exercises with these agents and learn how to use RoboNewbie and what the programming of robots is like. They can make their own experiences with different topics and algorithms by modifications and extensions.

It was a main goal of the project, to provide easily understandable concepts, methods and programs. There are no complicated structures, and all code is documented in detail. As a consequence, some more demanding concepts were replaced by simpler approaches (e.g. keyframe motions instead of inverse kinematics, approximated coordinates of observed objects etc.). Nevertheless, the clear structure of the project supports extensions for more challenging solutions if wanted.

A. Low Level Interface Functionalities

The framework includes interface functionalities on two levels. The lower one corresponds to the hardware-near functionalities of robots, while the higher one is concerned with more abstract control functionalities. Especially for the lower level, parts of the code of the team magmaOffenburg [8] were used by us as documented in our source files.

The hardware-near layer encapsulates the network protocol for interaction with SimSpark RCSS and it allows access to the simulated hardware entities corresponding to sensors and motors. The access is implemented by getter functions for perceptror values of different perceptrors which can be used similar to sensor signal queries of real robots. Related setter functions for effector values can be used for the control of actuators. Especially the low level interface functionalities for SimSpark RCSS are a hurdle for beginners and need time consuming work even for experienced users.

They concern tasks like network connection, synchronisation with the server, parsing of nested server messages, syntactical analysis of S-expressions, synthesis of agent messages with a lot of technical non-robotics details. The users of RoboNewbie need not to care about all this details, the framework offers ergonomic methods for the interaction with the simulated environment in an easily understandable way similar to the methods used by the operating systems of real robots. Users can learn to use these methods after a short training time (cf. the evaluation in Section VI).

The synchronization protocol was already described in Section IV. The user needs not to care about the communication, except the delays by the protocol and the duration of the cycles given by 20 msec. It is necessary to fetch a server message at each cycle and to send the agent message before the end of the cycle. The related control structures are already implemented in the examples and explained by the tutorial. Hence, if the calculations during one cycle do not exceed the cycle time, there will be no problem. The time needed depends of course on the used computer. The example agents run without problems even on less powerful machines.

The first example "Agent_BasicStructure" in the tutorial let the users start with an agent which already implements all low level communication. The agent simply rises an arm by setting related effector values. The user can experiment with other values and other effectors just to understand the basic structures.

B. Perception

The available perceptors were already listed in Section IV. All perceptor values can be queried by related getter methods using the perceptor names instead of the acronyms of the server messages. This allows a comfortable access to the perceptor data which corresponds to the access of sensor values by a related operating system of a real robot.

The necessary conversions from the nested server messages to the perceptor values are already implemented in the RoboNewbie framework. For that, the server message are parsed for the constituents of a tree like structure (again, thanks to the code of the team magmaOffenburg [8]). According to the analyzed acronyms in the expressions of the tree, the corresponding perceptor values are filled in by RoboNewbie.

The programs "Agent_TestPerceptorInput" and "Agent_TestLocalFieldView" illustrate the usage of the related getter methods and the perceptor values. The examples serve also as an illustration to the usage of the logger functions described below in Subsection E.

As an exercise of the tutorial, the user can implement an agent, which lifts the robots arm, when it senses another robot and moves the arm down, when it does not sense any robot. Which arm is lifted should depend on the side where the other robot is seen.

Special efforts are needed for the vision perceptor. It provides coordinates of all objects in the vision range of the camera. SimSpark RCCS in its common version does not communicate image data. Instead, the communicated

information can be understood as the result of basic image interpretation, it contains coordinates of the goal posts, the lines, the ball, and the body parts of robots.

The vision perceptor provides values by egocentric coordinates relatively to the camera in the centre of the head. Since the head may be turned and tilted, further calculations are necessary to get the coordinates of objects relatively to facing forwards. To get the coordinates relatively to the feet on the ground, it needs further calculations. Accurate calculations would need the inspection of the cinematic chain. The necessary data are available by the hinge joint perceptors. Further calculations including self localization are necessary for the transformation into global coordinates. RoboNewbie does not provide related programs following the intended "minimalistic" approach, because they would not be understandable by beginners without pre-knowledge about Robotics. But the implementation of related methods can serve as exercises during courses in Robotics.

As a simple substitute, we have decided to provide only approximations for the conversion from camera coordinates to facing forward coordinates. They are documented in the sources and easily to understand. Users can make experiments according to the accuracy and draw own conclusions on cinematic relations.

Visual information is provided by SimSpark RCSS only at each third cycle, and the robot would have to act blindly in between when there are no vision data available. Hence, the vision information should be stored for the following cycles. Moreover, the vision perceptor is limited by the camera view range of 120 degrees horizontally and vertically. The robot has to move its head to observe more objects in the world.

Again it is useful to store objects seen before in other directions. In general, such updating and memorizing of observations is maintained as belief of the robot in a so called world model. Updates may regard corrections according to robot motion, guesses for movements of invisible objects and integration of information communicated by other robots.

Again, a fully elaborated world model is far behind the scope of beginners. Hence, RoboNewbie provides a very simple version, where just the observed objects are stored in a simple form. The coordinates of those objects are referenced with respect to the facing forwards coordinates. Turnings of the head are already regarded by RoboNewbie, but only by the approximate calculations as described above. Other movements of the robot like turning or walking are not regarded. Time stamps indicate the last time of observing an object. The example "Agent_TestLocalFieldView" illustrates the perception features of RoboNewbie.

C. Motions

All intentional motions are performed by controlling the hinge joints (see Figure 4) by sending effector values (defining the speeds of motors) to SimSpark RCSS. Then the physics simulation engine calculates the effects of the commands regarding physical laws and updates the simulated world accordingly.

Simple motions like turning the head or rising the arms can be easily programmed by the users following the already mentioned examples. The motions can be controlled using the feedback of hinge joint perceptors. i.e. by sensor-actor coupling, where the delay of observing an action has to be regarded as described in Section IV. There is much room for own experiments of users.

More complicated motions like walking need coordinated movements of different joints. Users may learn about these problems after some trials. We have decided to provide keyframe motions in RoboNewbie because they are easily to understand and to design. The interpolation mechanism for keyframe motions in RoboNewbie realizes a linear interpolation - users may implement other interpolation methods like splines if they want. Keyframes are stored as text files which can be edited by any text processing system. Users can even design and change motions while using the programs as a black box.

these motions (the related text files). New motions need an integration as explained in the tutorial and the documentation.

According to simplicity, there are no concepts implemented for interruption of motions: Each motion is performed completely until its end, and there are no cyclic motions, e.g. for walking. Instead, continuous walking can be performed by subsequent calls of a two-step-walk.

The design of keyframe motions is supported by a graphical Motion Editor. It can be downloaded from the RoboNewbie Web page as well. It shows the postures of the robot for selected keyframes. Then the keyframes can be edited in two ways. In the graphical representation the posture can be kneaded into the desired posture with the mouse. Alternatively, each joint angle can be set to specified values which are immediately presented by the graphics. Transitions between keyframes can be defined with specific transition times resulting in a keyframe sequence as usual.

The program "Agent_KeyframeDeveloper" helps designing keyframes. A robot performs the motion of the actually edited keyframe file. After each change, the new motion is performed immediately. If the robot falls down, it stands up by itself.

The example "Agent_SimpleWalkToBall" illustrates the motion concepts. As an exercise of the tutorial, the users can change that program to implement obstacle avoidance (walk around the ball without touching it). They can use motions for walk, stop and turn. Additionally, the agent must be able to recognize the ball and to decide for the appropriate motion according to the ball position. Another exercise is the design of a new motion for kicking the ball. Users can furthermore do their own experiments e.g. with dancing robots.

In general, keyframe motions are useful for special motions like standing up, but they are not so well suited e.g. for walking. Walking is still a challenging problem in Robotics. The users of RoboNewbie will get some understanding about the task. Moreover, the framework is well suited as a basis for other implementations and for Machine Learning by more educated users. But according to our "minimalistic" approach, related implementations are not provided.

D. Control Cycle and Decision Making

The basic control cycle follows the classical centralistic deliberation approach, often denoted as the "sense-think-act-cycle", or by similar names. This corresponds closely to the cycle given by SimSpark RCSS: At first, sensations are provided to the agent, then the agent decides for appropriate plans and then it sends the related action commands back to the server.

To realize concepts of Embodied Robotics/AI it would be necessary to have local sensor actor coupling, distributed control, embodiment, situatedness, emergent behaviour etc. The real robot Nao as well as its simulated counterpart with the central control (i.e. our agent) is not primarily designed for such purposes. It is possible in principle to design sensor actor couplings and other behavioral concepts in the RoboNewbie framework. One might even split the agent into different "parallel" acting parts (implemented e.g. by threads) to simulate

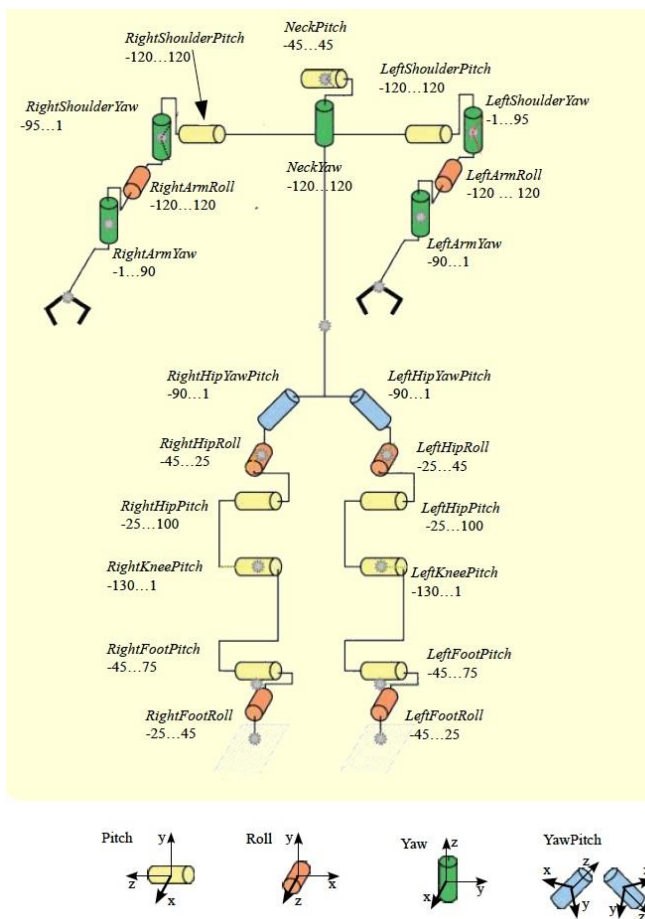


Figure 4: Joints of the robot. The range of the joint angles are given in degrees. Picture adapted from [5].

RoboNewbie contains a set of predefined keyframe motions for walking, turning, stand up and others. Users can change

distributed controls, but some synchronization is unavoidable by the server cycles of SimSpark RCSS.

At the same time, thinking in terms of the "sense-think-act-cycle" is quite natural for beginners because it reflects some causal dependencies. It provides an intuitive and easily maintainable structure in the design of robots. Therefore, the control cycle in RoboNewbie adopts the related terms for structuring the run-methods of the agents by cyclic calls of methods sense, think and act. The think-method is sometimes omitted in case of simpler ("reactive") agents.

The sense method is responsible for receiving and processing the perceptor data by the related RoboNewbie methods. The act method calls the transfer of the agent message with the effector commands. The think method between sense and act does the analysis of the perceptor data (e.g. a more elaborated world model) and the decision for plans and actions to be performed by the robot now and possibly in the future. The think method can of course be split into more dedicated deliberation methods which may be organized hierarchically if needed. All this can be worked out at the exercises during related courses. RoboNewbie provides just a simple example for illustration, the Agent_SimpleSoccer.

The Agent_SimpleSoccer is able to perform a very simple soccer play: As long as it is behind the ball and sees the opponent goal, it walks forward while pushing the ball with its feet. If the condition is not fulfilled, it turns around until it sees the ball, walks to the ball, turns around the ball until it sees the opponents goal, and then it starts walking towards the goal again. The decisions are made by a simple decision tree whenever the previous keyframe motion is completed (note that keyframe motions should not be interrupted).

Agent_SimpleSoccer can be improved in many ways. This is just what we want: Users can collect many ideas for improvements. They may concern better perception (e.g. by a ball model guiding the search), improved motions (like faster walk), new motions (like kick or dribble), better control (like path planning). It is possible to have more players on the soccer field such that players can cooperate (e.g. by positioning and passing). This gives room for simple contests during a course.

E. Logger

Runtime debugging of programs may be difficult because it affects synchronization with the server. Even simple debug messages printed on System.out may need too much time such that the agent cannot respond in time. It is possible to use the so-called sync mode which lets SimSpark RCSS wait until all agents have sent their messages (cf. the documentation). Alternatively, all debug messages can be collected by the program "Logger" of RoboNewbie. After the agent has finished, the collected messages are printed out. The usage is shown by the programs "Agent_TestPerceptorInput" and "Agent_TestLocalFieldView". Both programs provide also examples for the usage of the getter methods for perceptors.

VI. EXPERIENCES

The RoboNewbie framework was tested at different places. for introductory Robotics courses of about 30 hours during 5-8 days at Ohrid, Warsaw, Novi Sad, Rijeka, Sarajevo, and Plovdiv. 20 hours were planned for lectures, 10 hours for introduction and first usages of RoboNewbie. Additional 10-20 hours were used for further experiments by homework [9].

RoboNewbie served for illustrating experiments and for exercises in connection with the theoretical instructions. The participants of the courses learned to use RoboNewbie during short time and they programmed an improved soccer player at the end. The work with RoboNewbie was helpful to understand the theory. The final evaluation of the courses by the participants resulted in high marks. Especially the competitions with the improved soccer agent were motivating.

As RoboNewbie is intended for easy usage by beginners in Robotics, the requirements for the users are as minimal as possible, while the framework gives maximal support. For simplicity, approximations are used instead of complex calculations (e.g. simple offsets instead of linear algebra for determining the camera coordinates).

Since most of the courses had only a short duration, organisational issues were important for the success. We have asked the local organizers to prepare the technical resources accordingly. In the following, we describe some requirements in more detail.

A. Local Requirements for the Courses

Participants: Users are expected to have some programming skills in Java, such that they are able to understand and modify the agent programs. The programs are already prepared for usage under Netbeans, therefore the participants should be familiar with such tools. Users should be able to download and install programs from the web according to given instructions. Some physical and mathematical background is needed to understand the theoretical and practical issues of Robotics.

Participants should work in teams (as useful for programming exercises in general). Each team might consist of 3-5 participants, preferably mixed by different skills of its members. It helps for a smooth course if there are no big differences between the teams (e.g. each team should have at least one of the good programmers of the course, good mathematicians etc.).

Technical Resources: The participants should have their own computers where they can install and use the programs. Participants need access to the computers during the courses as well as for their homework. Hence, laptops are preferable. They are sufficient to run all the programs. Alternatively, participants may use computers in a lab (which have to be prepared accordingly if students are not allowed to install their own software).

The list of needed installations is given on the RoboNewbie webpage. Instructions for installation and functionality tests are

found there, too. If possible, students should get information before starting the course. They should be asked to install the programs by themselves and test if the programs can be started. If students cannot be asked before, an on-site test by some responsible person should be performed. It helps to save time during the courses if on-site problems with hardware or software are solved before. Nevertheless, if computers are ready, installation of programs needs only short time and can be done at the beginning of a course.

Organisational Issues: A good schedule is necessary for smooth courses. This includes early information (as far as possible) of participants as described above. Then the lectures and exercises are mixed appropriately. After a short overview about Robotics, participants start their first exercises as given by the "Quick Start Tutorial" (also found on the RoboNewbie web page). Later, more explanations are given as far as the theoretical lectures proceed. Thus, theoretical introductions to sensors can be connected to explanations of perceptor usage in RoboNewbie, introductions to motions are connected to the development of keyframes etc.

B. Competition

Our courses end with a competition, which serves as a motivation for the participants. The successful participation at the competition can also be a substitute for an examination if students need some certificate.

The competition is announced at the beginning of the course, and it should be performed by the teams. This helps for the integration inside the teams from the very beginning. The number of competing teams should be not more than 10 in order to make the contest not too long. This is also an argument to form teams if the number of participants is larger. The level of teams should be comparable for fairness reasons.

Until 2014, the task for the competition was an improvement of the program `Agent_SimpleSoccer` to get a better performance. It was up to the teams, what they wanted to improve. `Agent_SimpleSoccer` performs very poorly as described before. It needs about 10 minutes to find the ball and to push it into the goal. It was designed this way just to motivate the participants for improvements.

To make the competition a success (and a fun), it must be organized by strict and transparent rules. It should have a tight schedule to emphasize the aspects of sports. Therefore, each team has only one trial of only 3 minutes. The ranking of teams is determined by fastest scoring times. For teams who did not score, the ranking is given by minimal distances to the goal after the 3 minutes have elapsed.

At the competition, each team gives a short description of its efforts and expected results. This is also a possibility to check the engagement of each team member. Moreover, each participant can be asked to provide a written report of his/her individual efforts.

At the course in Plovdiv 2014, a student group implemented a very powerful kick which allowed scoring immediately. Thus this kind of task is considered to be solved finally (in times of internet, copying this solution could make

following competitions too simple). Thus, a new kind of competition with penalty kicks (attacker vs. goalkeeper) is tested in our recent courses.

C. Evaluation and Results

The participants of the courses were asked to give feedback on a prepared form at the end of the course. They could evaluate different aspects of the course and the framework. As the evaluation shows, the exercises with the simulated robots were motivating and helpful, the participants wanted to have more time for exercises and especially for own experiments.

As expected, the participants with less experience in Robotics gave higher marks related to motivation and help. The usage of the framework was intuitive. Interestingly, the participants with more experience in Java programming gave significantly higher rankings to the framework. The level of the exercises was considered as adequate, but for that the proportion of exercises was adapted by us accordingly.

As a unique observation, participants wanted to have more time for exercises than for lessons. This may have several reasons. The individual work load resulted in a bias for exercises: The participants had to fulfill given requirements, and many of them spent much time for preparing the final competition. Furthermore, the lectures tried to give a broad overview about the actual state of art in Robotics. There was not enough time to exercise on all these topics.

The "minimalistic approach" is useful especially for short courses and for introductions to longer courses. Later on, the disposability of non-minimalistic more sophisticated methods could be useful for higher level integrative tasks. It is impossible to let students implement all desirable algorithms in the limited time of a course. Joint activities of robots, for example, depend heavily on the available bodily skills and on the capabilities for interaction and coordination.

VII. CONCLUSION

The RoboNewbie framework can be used without special hardware. It simply needs a computer for simulation of the robot soccer scenario. The soccer scenario with humanoid robots is more complex than experiments by many hardware equipments. Nevertheless, RoboNewbie is easy to understand and to use after a short introduction. No special knowledge (except basic programming in Java) is required to start with own experiments, and while the users acquire more knowledge, they can work on more challenging tasks.

The practical evaluations have confirmed our expectations on the RoboNewbie project. Beginners in Robotics were able to use the framework after short introductions. They were able to program own methods in parallel to the theoretical concepts and methods provided by classes.

ACKNOWLEDGMENT

The first version of RoboNewbie was developed by Monika Domańska from the NaoTeam Humboldt [7]. We are thankful to the whole RoboCup community, especially to the developers

of SimSpark RCSS, to the team magmaOffenburg and to our team NaoTeam Humboldt, and especially to Yuan Xu.

REFERENCES

- [1] T. Padir, and S. Chernova (eds.), Special issue on robotics education. IEEE Transactions on Education, vol. 56, issue 1, 2013. <http://ieeexplore.ieee.org/xpl/tocresult.jsp?isnumber=6423944>. Visited at 19.1.2015.
- [2] Yuan Xu, From Simulation to Reality: Migration of Humanoid Robot Control. Dissertation Humboldt University Berlin, 2013.
- [3] RoboCup Web page. <http://www.robocup.org>. Visited at 19.1.2015.
- [4] Aldebaran Web presence <http://www.aldebaran-robotics.com/en/>. Visited at 19.1.2015.
- [5] SimSpark RCSS Wiki (Documentation of the Simulator). <http://simspark.sourceforge.net/wiki>. Visited at 19.1.2015.
- [6] R. Smith, Open Dynamic Engine User Guide, 2006. <http://www.ode.org>. Visited at 19.1.2015.
- [7] RoboNewbie. <http://www.naoteamhumboldt.de/projects/robonewbie/>. Visited at 19.1.2015.
- [8] Homepage Team magmaOffenburg. <http://robocup.hs-offenburg.de/>. Visited at 19.1.2015.
- [9] M. Domańska, H.D. Burkhard, RoboNewbie: A Framework for Experiments with Simulated Humanoid Robots. In M. Ivanović, L.C. Jain (eds.), E-Learning Paradigms and Applications, Agent-based Approach. Springer Series: Studies in Computational Intelligence, vol. 528, 2014, pp. 1-38.

Dynamic Frames-Based Generation of Web 2.0 Applications

Tihomir Orehovački

Faculty of Organization and Informatics
University of Zagreb
Pavlinska 2, 42 000 Varaždin, Croatia
tihomir.orehovacki@foi.hr

Ivan Magdalenić

Faculty of Organization and Informatics
University of Zagreb
Pavlinska 2, 42 000 Varaždin, Croatia
ivan.magdalenic@foi.hr

Danijel Radošević

Faculty of Organization and Informatics
University of Zagreb
Pavlinska 2, 42 000 Varaždin, Croatia
danijel.radosevic@foi.hr

Abstract—Frame Technology (FT) and Generative Programming (GP) are two widely accepted paradigms of software product lines development. While GP addresses the automatic generation of source code, FT advocates its adaptation to diverse reuse contexts. With an aim to utilize benefits of both approaches, this paper presents the SCT dynamic frames model that supports the automatic generation of Web 2.0 applications. The SCT model encompasses three essential components: Specification (S), which refers to application features, Configuration (C), which describes application development rules, and Template (T), which denotes application building blocks. Owing to its flexibility, readability, interactivity, and other object-oriented features, the Python scripting language was selected for the implementation of the generator. In order to demonstrate the appropriateness and usefulness of the proposed approach, an example that illustrates the generation of a Web 2.0 application for database management is provided.

Keywords—Web 2.0 Applications; Dynamic Frames; Generative Programming

I. INTRODUCTION

The term Web 2.0 refers to a second generation of web applications which enable users to interact with functionalities of their interfaces in a desktop-like fashion. Being dynamic in nature, Web 2.0 applications encourage users to create, share, publish, organize, and integrate a variety of artefacts thus contributing to the development of knowledge repositories. Given that Web 2.0 applications provide support for asynchronous and synchronous communication among users as well as collaboration on artefacts, they are commonly referred to as social web applications. According to Orehovački et al. [13], the most popular representatives of Web 2.0 applications are wikis, blogs, microblogs, social bookmarking sites, social networking sites, mashups, podcasting applications, e-portfolios, virtual worlds, online office suites, and knowledge management applications. Considering that evaluation presents indispensable part of every development process, recent research effort was focused on modelling their adoption [15], classification of quality in use metrics [42], measuring quality of

collaborative editors [19][20], evaluating the quality in use of mind mapping [17][18][20] and diagramming services [17] by means of both objective and subjective instruments, assessment of mashup tools [16], as well as evaluation of artefacts [21] which represent an outcome of their use.

From technical perspective, Web 2.0 applications are flexible services implemented in client-side Asynchronous JavaScript and XML (AJAX) frameworks. On the server side, scripting languages such as PHP, Perl, Python, Ruby, and JSP are used for delivering content from files and databases to the client. Despite the fact that Web 2.0 applications are widely used for both private and business purposes, there is a lack of comprehensive models and methodologies for their systematic development. Namely, the majority of current approaches deals with the model driven interface design (e.g. [11][12]), development (e.g. [3][10]), and code generation (e.g. [1][4]) of Rich Internet Applications (RIAs).

Considering the complementariness of different software development paradigms, a number of authors (e.g. [23], [30])

have merged two or more approaches into one thus yielding significant synergy effects. With an objective to achieve similar results in the context of Web 2.0 applications, we integrated concepts of frame technology (FT) and generative software development (GSD). Frame technology is a textual pre-processor which consists of two essential components: hierarchically organized code templates (frames), and a specification which contains particular features that can be adapted to different contexts [5]. On the other hand, generative software development supports mapping between a set of the features described by a domain specific language (DSL), and implementation components with all their possible combinations [2]. The aim of this paper is to illustrate appropriateness and usefulness of the use of SCT dynamic frames [7] in the generation of Web 2.0 applications.

The remainder of the paper is organized as follows. Overview of current research is provided in the second section. Features of the SCT generator model and generator design steps in the context of Web 2.0 applications are explained in the third section. An example how SCT generator can be employed for the purpose of developing Web 2.0 applications is illustrated in the fourth section. Concluding remarks and future research directions are offered in the last section.

II. BACKGROUND TO RESEARCH

The purpose of this section is to provide a brief review of two software development paradigms which constitute the theoretical background to the dynamic frames-based generation of Web 2.0 applications.

Software product line (SPL) denotes a group of software products that have a common set of features which meet stakeholders' needs [14]. Drawing on frame technology (FT), frame based software development (FBSD) is focused on design of generalized and adapted components. FT refers to a language independent textual pre-processor whose aim is development of systems which can be easily modified and consequently reused in a variety of contexts [5]. There are two essential elements which constitute frame technology: code templates structured in a hierarchy of modules known as frames and a specification that consist of particular features specified by the developer. In the context of software engineering, the aforementioned infrastructure represents a sound architecture for deriving SPLs [24]. Grossman and Mah [22] found that the employment of FT results in a decrease of expenses and time to market for large software development projects while in the same time contributes to the increase of reuse levels. These productivity enhancements motivated Jarzabek and Zhang [26] to introduce the meta-programming technique called XML-based Variant Configuration Language (XVCL) that drawing on Basset's frames [25] facilitates management of variability in SPLs. XVCL supports the decomposition of programs into generic and adaptable meta-components known as x-frames which as XML files represent domain knowledge in the form of SPL artefacts. An x-framework is a normalized layered hierarchical structure composed of x-frames that allows handling variants at different granularity levels. A configuration of a particular SPL member is managed by the topmost x-frame which is called the specification frame (SPC). Starting with the SPC call, the XVCL

processor goes through an x-framework, interprets XML tags in visited x-frames and by conducting necessary adaptations assembles components of specific SPL members. Taking into account advantages of XVCL with the respect to the reusability improvement, its concepts have been thoroughly evaluated in the context of databases [27], fault tolerant architectures [28], computer aided dispatch domain [29], etc.

The central role in generative software development (GSD) plays domain model which deals with mapping between problem space and solution space [2]. Problem space denotes a set of features of a SPL member that are described by means of the DSL. Implementation-based abstractions that constitute the specification of a SPL member are referred to as solution space. The mapping between the set forth spaces is carried out with the use of generator which calls a specification and eventually result in a corresponding implementation. Apart from XVCL, there are some other techniques that are also used for the purpose of generating software artefacts. One of them is GenVoca [31], a composition methodology meant for generating hierarchical SPL families. Fundamental features related to GenVoca are virtual machines, layers, realms, type equations, and a grammar. Virtual machines represent a set of methods, classes, and their objects that are employed for the implementation of SPL functionalities. An implementation of particular virtual machine is called layer. Realm is a set of layers that implement the same virtual machine. Each layer imports interface of the realm whose parameters it contains and exports the virtual machine of the realm it belongs to. Layer that imports and exports the same virtual machine is labeled as symmetric layer. The objective of layers is to encapsulate transformation that maps objects and operations between virtual machines. The structure composed of layers that are employed for modeling a particular software system is called a type equation. Realms together with their layer specify a grammar in which particular SPL member has a role of a sentence.

Current research related to the practical use of generators can be classified into several groups. The first group is focused on generating code snippets in a variety of programming languages that range from Python [39] and Java [6][8] to PHP [9]. The aim of the second group of generators is design of non-code artefacts such as graphical interface [38], programming assignments [36], text [37], and 3D scenes [41]. The last group is composed of generators which are implemented in scripting programming languages such as Open PROMOL [34] and CodeWorker [35]. While Open PROMOL deals with specifying program modifications of a target language, CodeWorker is meant for both parsing of arbitrary grammars and source code generation. The generator presented in this paper adds to the extant body of knowledge which deals with generation of code artefacts. Details on features of generation architecture that was employed for that purpose in the context of Web 2.0 applications are provided in the following section.

III. GENERATING THE WEB 2.0 APPLICATION

The SCT generator model is based on dynamic frames [7] and can be used in the generation of a wide variety of applications. The SCT generator model defines the generator of source code from three core elements: Specification (S),

Configuration (C) and Templates (T). Specification contains features of the generated application in the form of attribute-value pairs. Templates contain source code in a target programming language together with connections (replacing marks for the insertion of variable code parts). Configuration defines the connection rules between Specification and Templates. All three model elements together constitute the SCT frame.

A particular SCT frame produces source code that could be either stored in a specific data file or included in another SCT frame. The basic idea of the generation process is shown in Figure 1. The initial SCT frame contains the initial source code template that includes connections. The generator of the source code creates a new SCT frame for each connection. While the source code of SCT frames located deeper in the hierarchy is included as the integral part of its superior SCT frame, the source code of the initial SCT frame is stored in a data file.

Since an average application contains more data files, the SCT model implies the existence of a Handler. It represents a part of the SCT source code generator that aims to make the generator scalable in a way that it can produce more pieces of program code (e.g. program files) from the same set of Specification, Configuration and Templates. The SCT dynamic frames model enables the generation of various program units (e.g. files, classes, functions etc.) from the same Specification. Moreover, it enables the generation of code in a variety of programming languages (e.g. JavaScript, PHP, XML, Python, Java, etc.) and is consequently suitable for the generation of Web 2.0 applications. The generated code can be stored in program files for later execution as well as in variables for immediate execution [32] [32].

Web 2.0 applications are specific since they use different technologies in an integrated manner. The flexibility of the SCT generator enables implementation of several technologies in the same application. This is achieved with cautious design of Specification, Templates and Configuration of the SCT generator.

Model which reflects development process of Web 2.0 application by means of SCT generator is illustrated in Figure 2. The first step is to identify Web 2.0 services and build one or several prototypes for each service. Based on experience obtained during development of prototypes, a set of templates is developed for each service. Those templates are input into SCT generator. Different applications have different set of services which is defined in Application specification. Each application has list of web services and other important data listed in its Application specification. How templates are combined together, based on Application specification, is defined in Application configuration.

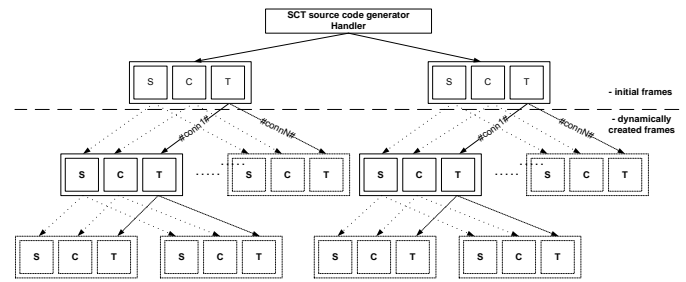


Fig. 1. The generating process

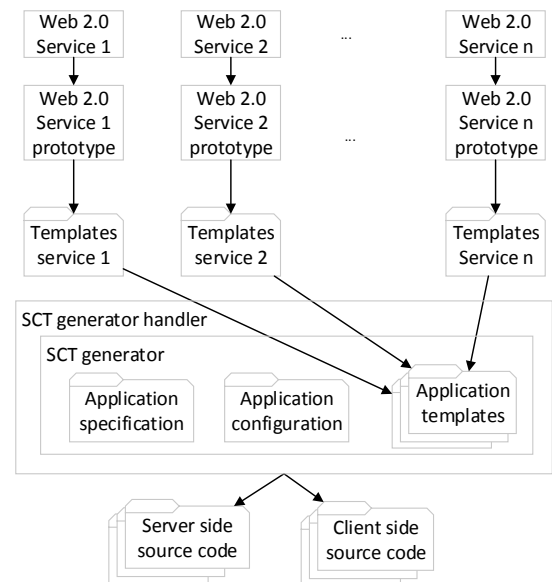


Fig. 2. Model of building Web 2.0 applications using the SCT generator

The SCT Handler generates more data files which contain source code and implement both server and client side of Web 2.0 application. Since Web 2.0 application employs different technologies, it is a challenge to make such templates that can be easily manageable and reusable. In that respect, the SCT generator model offers management of the whole set of code templates via relatively small Configuration.

The process of building new generators begins with application prototype that is decomposed into SCT model elements through several steps. The SCT generator applies these elements in automatic assembling of different application variants. Steps in the design of a generator of Web 2.0 application are as follows [40]:

0. *Prerequisite.* The prerequisite for building the SCT generator is the application prototype in a form of a source code.

1. *Selection of new main templates and output types.* The main templates are specified in the initial part of Configuration and define the type of code to be generated, e.g.:


```
#1#,,index.template - entry HTML page
```

2. *Creating of Specification.* Specification consists of attributes and their values. The hierarchy of attributes is specified by '+' sign, e.g.:

```
field_combo:id_course
+field_display:Course - subordinated attribute
```

3. *Delineation of variable program parts.* Variable program parts depend on Specification, so they will be later replaced with connections.

4. *Flexibilization of prototype.* Variable program parts are being replaced by connections (in #-es).

5. *Adding new rule to Configuration.* The configuration rule specifies all three elements of the SCT model: connection, specification attribute and used code template, respectively e.g.:

```
#links#,title,links.template
```

6. *Building of code templates* that are main constituent artefacts of generated applications.

7. *Generating, testing and adjusting in a generative development process.*

An outcome of the development process that begins with an application prototype is a generator that can be used in automatic design of different application variants. Obtained SCT model elements (Specification, Configuration and Templates) can be used in the further development of generators as well as applications.

As shown in Figure 3, the development process of particular Web 2.0 application can be illustrated with spiral model that was originally proposed by Boehm [43].

IV. THE EXAMPLE OF GENERATING

The example¹ includes a SCT based generator, implemented in Python, together with generated Web 2.0 application (also in Python; Ajax was used for user interface and PostgreSQL for database implementation).

Specification of the given instance contains three output types:

```
OUTPUT:out1 - used for index page
OUTPUT:output - CGI scripts (Python)
OUTPUT:output_html - HTML forms
```

Each output type refers to one or more output files that will be generated. For instance, there are two files that are going to be generated from the following specification group:

```
output:output/students.cgi - CGI script
output_html:output/students_form.html - HTML form
table:ajax_students - database table
connection:exams - link to another DB table
+connection_field:student_id - subordinated attributes
+connection_display:Exams to link
title:students - group name
+title_display:Students - text to be displayed
primary_key:student_id - DB table primary key
field_number:student_id - table attribute+type
+field_display:Student id - text to be displayed
```

¹ The example is available at gpmf.foi.hr/SCT_Python_Ajax

```
field_text:surname_name
+field_display:Surname and name . . .
field_number:year_of_enrollment
+field_display:Year of enrollment . . .
field_number:year_of_study
+field_display:Year of study
```

Configuration contains rules for assembling software from Specification and Templates. The initial part of Configuration specifies the initial code templates that correspond to output types from Specification:

```
#1#,,index.template - index page
#2#,,script.template - CGI scripts
#3#,,form.template - HTML forms
```

Other lines of Configuration contain two- or three-element groups, e.g.:

```
#table#,table - link, attribute
#title_field#,title,title.template - link, attribute,
template
```

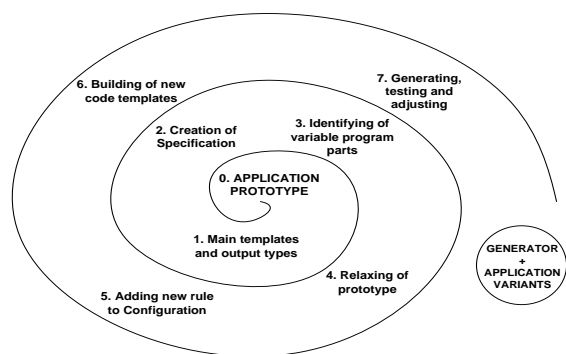


Fig. 3. Spiral generator/application development [40]

The two-element group specifies direct replacement of the position of variable in Templates with value of attribute from Specification. The three-element group specifies that code template has to be used as many times as it occurs in Specification. Each code template employs connections (usually words in #-es) in order to specify variable parts that are going to be generated. The following example of a template is used in the generation of input/edit forms:

```
<form id="myForm" action="" method="POST" >
<input type="hidden" name="action" value="!operation!">
#fields_on_form#
<td><input type='submit' name="Submit"
value='Send'onclick="Perform2('#fields_getelementbyid#',
'#title#.cgi?action=!operation!','R0')"></td>
<td>&nbsp;</td>
</form>
```

The example utilizes Ajax to route the output of the CGI script to a particular HTML element marked by id (here: 'R0'). As shown in Figure 4, this feature enables editing of particular row in database table, without the need of refreshing the whole web page.

The aim of the SCT generator model is to achieve high reusability of features (attributes with their values) defined in Specification. These features can be distributed through

connections on many different places in diverse code templates, as shown in Table 1.

TABLE I. DISTRIBUTION OF SPECIFICATION FEATURES IN THE EXAMPLE APPLICATION

Attribute	Total number of occurrences in Specification	Total number of occurrences in generated code	Number of files where the value occurs
application	1	2	1
table	3	90	3
title	3	52	3
title_display	3	15	3
field_number	6	397	3
field_combo	2	70	1

features specification at higher level of abstraction. The second one is simplified application update which results from definition of application in a higher abstraction language used by SCT generator. Inclusion of new features in application is performed by adding new definitions in application specification. The third one is the customization of application to the specific needs of particular user. Considering that network effects, perpetual beta, and lightweight user interfaces, respectively are essential design patterns of Web 2.0 applications, the proposed approach supports the user-centered development of software product line members.

Our future work will be focused on the employment of dynamic frames based generators in the development of some specific types of Web 2.0 applications such as mashups. More specifically, our research efforts will deal with interplay of different generator implementations and novel web technologies.

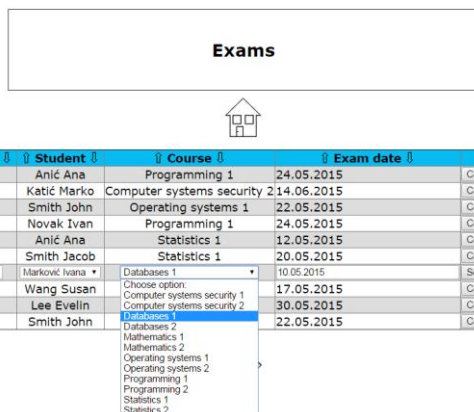


Fig. 4. Editing the particular row in database table

The multi-distribution of specification features could be used in application updating. This can be performed by changing the Specification, which enables new features of applications inside the problem domain proposed by Configuration. Any modification in Templates changes the way Specification attribute values are used, including the programming language. Any update of Configuration changes the way the generator builds the program code, respectively. The introduction of a new line in Configuration could enable the use of a new Specification attribute and a new code template. The purpose of the set forth is to avoid any later modifications of the generated code.

V. CONCLUDING REMARKS

This paper illustrated the use of dynamic frames generator model in the development of Web 2.0 applications. There are numerous benefits of the proposed approach. The first one is the improvement of the development process productivity that is the outcome of reusability of program artefacts. The set forth productivity reflects in terms of enhanced efficiency in development of software product lines, as well as facilitated

REFERENCES

- [1] A. Bozzon and S. Comai, "Conceptual Modeling and Code Generation for Rich Internet Applications", Proceedings of the 6th International Conference on Web Engineering, pp. 13-18, 2006.
- [2] K. Czarneci, "Overview of generative software development", in Unconventional Programming Paradigms, Lecture Notes in Computer Science, vol. 3566, J.-P. Banâtre, P. Fradet, J.-L. Giavitto and O. Michel, Eds. Le Mont Saint Michel: Springer, 2004, pp. 326-341.
- [3] J.M. Hermida, S. Meliá, A. Montoyo and J. Gómez, "Developing Rich Internet Applications as Social Sites on the Semantic Web: A Model-Driven Approach", International Journal of Systems and Service-Oriented Engineering, vol. 2, no. 4, pp. 21-41, 2011.
- [4] M. Linaje, J.C. Preciado, R. Morales-Chaparro, R. Rodríguez-Echeverría and F. Sánchez-Figueroa, "Automatic Generation of RIAs Using RUX-Tool and Webratio", in Web Engineering, Lecture Notes in Computer Science, vol. 5648, M. Gaedke, M. Grossniklaus and O. Díaz, Eds. San Sebastian: Springer, 2009, pp. 501-504.
- [5] N. Loughran, A. Rashid, W. Zhang and S. Jarzabek, "Supporting Product Line Evolution with Framed Aspects", in Workshop on Aspects, Components and Patterns for Infrastructure Software, 2004, <http://www.comp.lancs.ac.uk/computing/aod/papers/SPL_ACP4IS2004.pdf>.
- [6] I. Magdalenic, D. Radošević and Z. Skočir, "Dynamic Generation of Web Services for Data Retrieval Using Ontology", Informatica, vol. 20, no. 3, pp. 397-416, 2009.
- [7] D. Radošević and I. Magdalenic, "Source Code Generator Based on Dynamic Frames", Journal of Information and Organizational Sciences, vol. 35, no. 1, pp. 73-91, 2011.
- [8] D. Radošević, M. Konecki and T. Orehovački, "Java Applications Development Based on Component and Metacomponent Approach", Journal of Information and Organizational Sciences, vol. 32, no. 2, pp. 137-147, 2008.
- [9] D. Radošević, T. Orehovački and M. Konecki, "PHP Scripts Generator for Remote Database Administration based on C++ Generative Objects", Proceedings of the 30th MIPRO Jubilee International Convention on Intelligent Systems, pp. 167-172, 2007.
- [10] B. Steam, "XULRunner: A New Approach for Developing Rich Internet Applications", Internet Computing, vol. 11, no. 3, pp. 67-73, 2007.
- [11] M. Urbietta, G. Rossi, J. Ginzburg and D. Schwabe, "Designing the Interface of Rich Internet Applications", Proceedings of the 5th Latin American Web Conference (LA-WEB), pp. 144-153, 2007.

- [12] F. Valverde and O. Pastor, "Applying Interaction Patterns: Towards a Model-Driven Approach for Rich Internet Applications Development", Proceedings of Workshop on Web-oriented Software Technology (IWWOST), pp. 13-18, 2008.
- [13] T. Orehovački, G. Bubaš and A. Kovačić, "Taxonomy of Web 2.0 Applications with Educational Potential", in Transformation in Teaching: Social Media Strategies in Higher Education, C. Cheal, J. Coughlin and S. Moore, Eds. Santa Rosa: Informing Science Press, 2012, pp. 43-72.
- [14] P. Clements and L. Northrop, Software product lines: Practices and patterns. Boston: Addison-Wesley, 2002.
- [15] T. Orehovački and S. Babić, "Predicting Students' Continuance Intention Related to the Use of Collaborative Web 2.0 Applications", Proceedings of the 23rd International Conference on Information Systems Development (ISD), pp. 112-122, 2014.
- [16] T. Orehovački and T. Granollers, "Subjective and Objective Assessment of Mashup Tools", in Design, User Experience, and Usability - Theories, Methods, and Tools for Designing the User Experience, Lecture Notes in Computer Science, vol. 8517, A. Marcus, Ed. Heraklion: Springer, 2014, pp. 340-351.
- [17] T. Orehovački, A. Granić and D. Kermek, "Evaluating the Perceived and Estimated Quality in Use of Web 2.0 Applications", Journal of Systems and Software, vol. 86, no. 12, pp. 3039-3059, 2013.
- [18] T. Orehovački, A. Granić and D. Kermek, "Exploring the Quality in Use of Web 2.0 Applications: The Case of Mind Mapping Services", in Current Trends in Web Engineering, Lecture Notes in Computer Science, vol. 7059, A. Harth and N. Koch, Eds. Paphos: Springer, 2011, pp. 266-277.
- [19] T. Orehovački, "Perceived Quality of Cloud Based Applications for Collaborative Writing", in Information Systems Development – Business Systems and Services: Modeling and Development, J. Pokorny, V. Repa, K. Richta, W. Wojtkowski, H. Linger, C. Barry and M. Lang, Eds. Prague: Springer, 2010, pp. 575-586.
- [20] T. Orehovački, S. Babić and M. Jadrić, "Exploring the Validity of an Instrument to Measure the Perceived Quality in Use of Web 2.0 Applications with Educational Potential", in Learning and Collaboration Technologies - Designing and Developing Novel Learning Experiences, Lecture Notes in Computer Science, vol. 8523, P. Zaphiris, and A. Ioannou, Eds. Heraklion: Springer, 2014, pp. 192-203.
- [21] T. Orehovački and N. Žajdela Hrustek, "Development and Validation of an Instrument to Measure the Usability of Educational Artifacts Created with Web 2.0 Applications", in Design, User Experience, and Usability - Design Philosophy, Methods, and Tools, Lecture Notes in Computer Science, vol. 8012, A. Marcus, Ed. Las Vegas: Springer, 2013, pp. 369-378.
- [22] I. Grossman and M. Mah, "Independent research study of software reuse using frame technology", Technical Report, QSM Associates, 1994.
- [23] L. Fuentes, C. Nebreira and P. Sánchez, "Feature-oriented model-driven software product lines: the TENTE approach", Proceedings of the forum of the 21st international conference on advanced information systems (CAiSE), pp. 67-72, 2009.
- [24] P.G. Bassett, "The case for frame-based software engineering", IEEE Software, vol. 24, no. 4, pp. 90-99, 2007.
- [25] P.G. Bassett and E. Yourdon, Framing software reuse – lessons from real world. Upper Saddle River: Prentice Hall, 1997.
- [26] S. Jarzabek and H. Zhang, "XML-based method and tool for handling variant requirements in domain models", Proceedings of the 5th IEEE international symposium on requirements engineering, pp. 166-173, 2001.
- [27] S. Guo, L. Tang and W. Xu, "XVCL – an annotative approach to feature-oriented programming", Proceedings of the 2010 international conference on computational intelligence and software engineering, pp. 1-5, 2010.
- [28] L. Yuan, J. Song Dong and J. Sun, "Modeling and customization of fault tolerant architecture using object-Z/XVCL", Proceedings of the 13th Asia Pacific software engineering conference, pp. 209-216, 2006.
- [29] H. Zhang and S. Jarzabek, "XVCL: a mechanism for handling variants in software product lines", Science of Computer Programming, vol. 53, no. 3, pp. 381-407, 2004.
- [30] I. Groher and M. Voelter, "Aspect-oriented model-driven software product line engineering", in Transactions on Aspect-Oriented Software Development VI, Lecture Notes in Computer Science, vol. 5560, S. Katz, H. Ossher, R. France and J-M. Jézéquel, Eds. Heidelberg: Springer, 2009, pp. 111-152.
- [31] D. Batory, V. Singhal, J. Thomas, S. Dasari, B. Geraci and M. Sirkin, "The GenVoca model of software-system generators", IEEE Software, vol. 11, no. 5, pp. 89-94, 1994.
- [32] R. Fabac, D. Radošević and I. Magdalenić, "Autogenerator-Based Modelling Framework for Development of Strategic Games Simulations: Rational Pigs Game Extended", The Scientific World Journal, 2014, <dx.doi.org/10.1155/2014/158679>.
- [33] I. Magdalenić, D. Radošević and T. Orehovački, "Autogenerator: Generation and Execution of Programming Code on Demand", Expert Systems with Applications, vol. 40, no. 8, pp. 2845-2857, 2013.
- [34] V. Štūkys and R. Damaševičius, "Scripting language open PROMOL and its processor", Informatica, vol. 11, no. 1, 71-86, 2000.
- [35] C. Lemaire, "CodeWorker parsing tool and code generator – user's guide & reference manual, release 4.5.4.", 2010, <<http://www.codeworker.org/CodeWorker.pdf>>.
- [36] D. Radošević, T. Orehovački and Z. Stapić, "Automatic on-line generation of student's exercises in teaching programming", Proceedings of the 21st Central European conference on information and intelligent systems, pp. 87-93, 2010.
- [37] J. Müller and U.W. Eisenacker, "The applicability of common generative techniques for textual non-code artifact generation. In Proceedings of the workshop on modularization, composition, and generative techniques for product line engineering, 2008, <<http://www.infosun.fim.uni-passau.de/spl/apel/McGPLE2008/papers/Paper8.pdf>>.
- [38] M. Schlee and J. Vanderdonck, "Generative programming of graphical user interfaces", Proceedings of the working conference on advanced visual interfaces, pp. 403-406, 2004.
- [39] D. Radošević and I. Magdalenić, "Python implementation of source code generator based on dynamic frames", Proceedings of the 34th International Convention on Information and Communication Technology, Electronics and Microelectronics, pp. 369-374, 2011.
- [40] D. Radošević, I. Magdalenić and T. Orehovački, "Building Process of SCT Generators", Proceedings of the 36th International Convention on Information and Communication Technology, Electronics and Microelectronics, pp. 1037-1042, 2013.
- [41] A. Kvesić, D. Radošević and T. Orehovački, "Using SCT Generator and Unity in Automatic Generation of 3D Scenes and Applications", Proceedings of the 25th Central European Conference on Information and Intelligent Systems, pp. 312-317, 2014.
- [42] T. Orehovački, D. Kermek and A. Granić, "Examining the Quality in Use of Web 2.0 Applications: A Three-Dimensional Framework", Communications in Computer and Information Science, vol. 373, pp. 149-153, 2013.
- [43] B. W. Boehm. "A Spiral Model of Software Development and Enhancement". Computer, vol. 21, no. 5, pp. 61-72, 1988.

Unsupervised Classification of Mobile Device Images

Jocelin Rosales Corripio, Ana Lucila Sandoval Orozco, Luis Javier García Villalba

Group of Analysis, Security and Systems (GASS)
Department of Software Engineering and Artificial Intelligence (DISIA)
Faculty of Information Technology and Computer Science, Office 431
Universidad Complutense de Madrid (UCM)
Calle Profesor José García Santesmases, 9
Ciudad Universitaria, 28040 Madrid, Spain
Email: jocelinr@ucm.es, {asandoval, javiergv}@fdi.ucm.es

Abstract— As mobile devices are seeing widespread usage in the everyday life, the images from mobile devices can be used as evidence in legal purposes. Accordingly, the identification of mobile devices images are of significant interest in digital forensics. In this paper, we propose a method to determine the mobile devices camera source based on the grouping or clustering of images according to their source acquisition. Our clustering technique does not involve a priori knowledge of the number of images or devices to be identified or training data for a future classification stage. The proposal combines of hierarchical and flat clustering and the use of sensor pattern noise. Experimental results show that our approach is very promising for identifying mobile devices source.

Keywords— *Image Clustering; Image Forensics Analysis; PRNU; Sensor Pattern Noise*

I. INTRODUCTION

Nowadays, even suffering the impact of global financial crisis, the sales of mobile devices such as cell phones, smartphones or tablets, is still increasing. About 78.1% of mobile phones sold in 2010 have an integrated camera [1]. Integrated cameras in mobile devices outnumber traditional Digital Still Camera (DSCs). The sales of cameras integrated into mobile devices in 2013 exceeded 1800 million units. Similarly, there are predictions that the DSCs will disappear in favour of integrated mobile devices [2], since the quality of these cameras is growing at an unstoppable rate. Also, the emergence of cameras in mobile devices should not only be measured in sales figures, as in our daily life it is common to see how people use photographs from these devices for a variety of situations – personal life, news, legal evidence, software applications and so on. Therefore, forensic analysis of such images is particularly important in criminal investigations.

The image source acquisition identification and malicious tampering detection are of significant interest in digital image forensic analysis. This work focuses on the first branch. Also, since mobile device cameras have some characteristics that make them different from the rest, this work focuses on images from this type of devices. The source acquisition identification has closed scenarios and open scenarios approaches regarding. A closed scenario is one in which the image source identification is performed on a specific and known beforehand set of cameras. In closed scenario approach normally use to train and predict process in order to classify like Support Vector Machine (SVM) classifier. Instead, in open scenarios the forensic analyst does not know a priori the camera set to

which images whose source identification will be identified belong.

In this paper, we propose a method that utilizes the hierarchical and flat clustering to image source identification in open scenarios. The objective of this approach is to group the different images into disjoint sets in which all their images belong to the same device. This approach is very close to real-life situations, since in many cases the set of cameras to which a set of images may belong is completely unknown to the analyst. In addition, it is virtually impossible to have a set of images to train a classifier with all mobile device cameras existing in the world. In this case, being able to group images into sets that belong to the same device is very useful, as this can provide very valuable and in some cases conclusive information to judicial investigators. The remainder of this paper is organized as follows. Section 2 briefly presents previous work related to forensic techniques for mobile device image source acquisition identification. The proposed technique is presented in section 3. The experiments and their results are presented in section 4. Finally, in section 5 the conclusions drawn from this work are presented.

A. Image Formation in Digital Cameras

The first step is to understand and create image processing forensic algorithms is to thoroughly know the process of image acquisition in digital cameras. Fig. 1 summarizes this process.

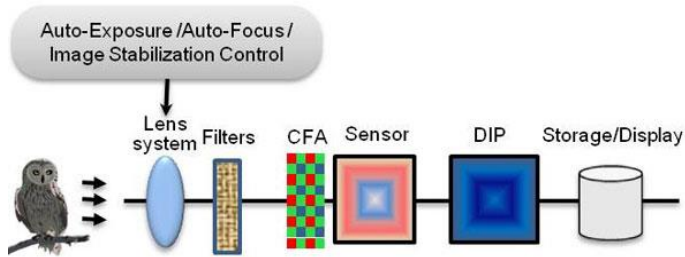


Fig. 1. Image acquisition process in digital cameras [3]

First, the lens system captures light from the scene by controlling the exposure, focus, and image stabilization. Next, the light passes through a set of filters that improve the visual quality of the image, and then the light gets to the image sensor called Color Filter Array (CFA); this is an array of light sensitive elements called pixels. Note that the choice of the CFA can influence the sharpness and the final appearance of the image since there are different CFA patterns.

The most commonly used model is the Green-Red-Green-Blue (GRGB) Bayer pattern; other models are: Red-Green-Blue-Emerald (RGBE), Cyan-Yellow-Yellow-Magenta (CYYM), Cyan-Yellow-Green-Magenta (CYGM) or Red-Green-Blue-White (RGBW). The incident light on the colored filters gets to a sensor which is responsible for generating an analogue signal proportional to the intensity of received light, keeping these values in an internal array.

There are currently two types of sensor technologies that meet this latter purpose in digital cameras: CCD (Charge Coupled Device) and CMOS (Complementary Metal Oxide Semiconductor). Both types of sensors essentially consist of Metal Oxide Semiconductors (MOS) and they work in a similar way, although the key difference is in the way in which pixels are scanned and the way in which the reading of the charges is carried out. CCD sensors need an additional chip to process the sensor's output information; this causes the manufacture of devices to be more costly and the sensors to be bigger. In contrast, CMOS sensors have independent active pixels, as they themselves perform the digitalization, offering speed and reducing the size and cost of the systems that make up a digital camera. Another difference between these two types of sensors is that the pixels in a CCD array capture light simultaneously, which promotes a more uniform output. CMOS sensors generally perform the reading as progressive scan (avoiding the blooming effect). CCD sensors are far superior to the CMOS in terms of noise and dynamic range; on the other hand, CMOS sensors are more sensitive to light and behave better in low light conditions. Early CMOS sensors were somewhat worse than CCDs, but nowadays this has been practically corrected.

The CCD technology has reached its limit and it is now when CMOS is being developed and its weaknesses are being overcome, so much that the majority of smartphones contain CMOS sensors. Signals stored by the CCD/CMOS sensor are then converted into a digital signal and transmitted to the image processor, once the image processor receives the digital signal it eliminates noise and other introduced anomalies. Some other

processes applied to the signal are color interpolation, gamma correction, and color correction.

II. PREVIOUS WORKS IN IMAGE FORENSIC ANALYSIS

Most research on image source acquisition identification focuses on traditional digital cameras or DSC; most of these techniques are not valid for mobile device images. In [4] an overview of this research can be seen.

For any type of image classification, either in open or closed scenarios, it is necessary to obtain certain features that allow classification techniques to perform their task. According to [3], four groups of techniques can be established for this purpose: based on lens aberration, based on the CFA matrix interpolation, based on the sensor imperfections and based on the use of image features. Within the latter group a subdivision can be made based on color features, quality features, and wavelet domain statistics. This work uses techniques based on sensor imperfections, particularly those based on the sensor pattern noise (SPN). The main components of image noise are the Fixed Pattern Noise (FPN) and the Photo Response Non Uniformity (PRNU). There are several sources of imperfections and noise introduced at different stages of the creating pipeline of an image in a digital camera. Even if a uniform and fully lighted picture is taken it is possible to see small changes in the intensity between pixels. This is due to the shot noise is random and, in large part, the pattern noise is deterministic and is kept approximately equal if several pictures of the same scene are taken.

The analysis of clusters, or clustering, aims to group a collection of objects into representative classes called clusters, without a priori information, in such a way that the objects belonging to each cluster keep a greater similarity to objects from other clusters. Image grouping can be performed using supervised or unsupervised learning techniques. In the first case it is essential to know the device information a priori, i.e., it is clearly identified with the classification in closed scenarios which requires a training stage with the features extracted from the images and a second classification stage in accordance with the previous result. However, in a real case it may be difficult to have the camera in question or a set of photographs taken by it to carry out training, hence the need for unsupervised learning techniques, which directly correspond to open scenarios.

Traditional clustering has been known to be an unsupervised learning technique; however, there are some cases of supervised clustering where it is possible to apply an anterior or posterior approach to improve the grouping itself. This is to prevent that elements of different classes are in the same cluster, which requires having a priori knowledge of the data set. This issue is addressed in [6], although it is worth mentioning that this article is focused on the use of unsupervised techniques.

In order to determine the similarity between objects belonging to the same cluster, there are distance measures such as Euclidean distance, Manhattan distance, and Chebychev distance, among others. Alternatively, it is possible to use

similarity functions $S(X_i, X_j)$ which compare two vectors X_i and X_j symmetrically, i.e., $S(X_i, X_j) = S(X_j, X_i)$. These functions reach their highest values as X_i and X_j are more similar. One of the most commonly used measures in image source identification is normalized correlation [7][8][17] defined in equation 1.

$$corr(X_i, X_j) = \frac{(X_i - \bar{X}_i) \odot (X_j - \bar{X}_j)}{\|X_i - \bar{X}_i\| \|X_j - \bar{X}_j\|} \quad (1)$$

Where \bar{X}_i and \bar{X}_j represent the mean vector, $X_i \odot \bar{X}_j$ is the scalar product of two vectors and $\|X_i\|$ is the L_2 norm of X_i .

According to the clustering algorithms classification proposed in [9], we find the hierarchical methods whose purpose is to achieve a structure called dendrogram which represents the grouping of objects according to their levels of similarity. This grouping can be done in different ways: agglomerative or divisive. Agglomerative grouping initially considers each object as a separate class until iteratively grouping all the objects in a single class. Divisive clustering is based on the idea of starting from a single class until managing to separate all objects into individual classes. There are also partitioning algorithms, wherein starting a partition, the algorithm takes care of moving objects from one cluster to another to minimize certain error criterion. Within this category, the most famous method is k-means; however, most of these methods require knowing in advance the number of clusters, which is why they are not widely used in forensic image analysis. Finally, there are other clustering algorithms such as: [10] which produces clusters by means of graphs, [11] based on the density where the points within a cluster are given by a certain probability function, clusters based on models such as decision trees [12] or neural networks [13] and clustering with soft-computing methods such as fuzzy clustering [14], evolutionary clustering methods and simulated annealing clustering [15].

There are previous works on image grouping by unsupervised methods; all of them consider SPN as the most reliable criterion for representing a device's digital footprint, hence the PRNU is used specifically as a footprint and normalized correlation as a similarity measure to achieve image grouping by device.

[16] uses a classification technique with unsupervised learning where grouping is achieved by graph maximization. Clustering is performed from not-oriented graph with weights, starting with an affinity matrix where the connection weights between vertices is the correlation value between each SPN, starting with a random node. In each iteration, the remaining nodes are connected and the nodes closest to the central one are chosen, obtaining a new affinity matrix in each step; the algorithm stops when the number of closest nodes is less than a k parameter. Subsequently, the graph is partitioned to the point where similarity in a set is maximum and minimum with respect to other sets.

In [8] clusters are performed using Markov random fields. A clustering algorithm based on matrix containing all the correlations between the SPN of several cameras is proposed.

In each iteration the algorithm groups within classes the most similar SPNs making use of the local features of Markov random fields and assigns a new class label to each SPN maximizing a probability function, the criterion to stop the algorithm is satisfied when there are no label changes after a certain number of iterations.

The algorithm proposed in [17] and on which this research is based uses hierarchical clustering to group images. Prior to the clustering algorithm, the authors apply a function for sensor noise improvement, which strengthens the lower components and attenuates the high components in the wavelet domain in order to remove the scene details in it. With a similarity matrix containing all the correlations between different SPNs and taking as a starting point each image as a single cluster, the clustering algorithm groups the two clusters with the highest correlation value forming a single cluster and updates the matrix with a new row and column that replace the rows and columns of the grouped clusters. The link criterion chosen to mix two clusters was average linkage. In each iteration of the algorithm, cluster status at that time is stored on a partition and the global silhouette coefficient is calculated. At the end of the algorithm the partition whose silhouette coefficient value is the lowest is chosen, the number of clusters at that point should correspond to the number of devices that exist initially, as well as the content of each cluster to the SPN for each device. The authors carry out a training stage with the described algorithm and a classification stage for the remaining images, for this it is sufficient to obtain the average of the SPNs for each cluster and compare them against the remaining images, the image will be classified within the cluster whose correlation is highest.

III. TECHNIQUE DESCRIPTION

The proposed unsupervised clustering algorithm is based on the one proposed in [17]. It is a combination of a hierarchical clustering, and a flat clustering. That is, despite forming a dendrogram structure with each iteration of the algorithm, at the end the clusters are taken as unrelated entities since each of them must correspond to a specific device.

Prior to performing the clustering, it is necessary to obtain sensor pattern noises of the image set I using the extraction algorithm and the parameter of noise suppression $s_0 = 5$ proposed in [5]. Equation 2 shows this calculation.

$$n^{(i)} = I^{(i)} - F(I^{(i)}) \quad (2)$$

Where $i = 1, \dots, N$, N is the number of images, $n^{(i)}$ is the noise pattern of each image i , $I^{(i)}$ is the image with sensor noise of each image i and F is the noise removal filter based on wavelet transform. For this, the algorithm developed by Goljan et al. in [18] was used. No noise improvement algorithm, such as those proposed by [8] and [17], has been used in our proposal. The Wiener filter in the frequency domain is sufficient to remove most of the scene details that are present when extracting the SPN.

For each of the N noises n_1, \dots, n_N the correlation value is obtained using equation 1 and this generates a similarity matrix H of $N \times N$. This matrix is symmetric and consists of ones in

its main diagonal (since the correlation of noise with itself is 1). Once the matrix has been generated it will not be necessary to recalculate the correlations between noises along the clustering algorithm, saving time and processing power.

The selected hierarchical clustering algorithm involves finding within the H matrix the noise pair k and l with a highest correlation value. It is worth mentioning that the correlation values in the main diagonal are not taken into account. Then the rows and columns k and l are deleted and both a new row and a new column are added to the matrix. These new row and column values are the result of a linkage criterion. The function chosen for this work was the average linkage method since its results are more satisfactory than with other linkage methods such as single linkage or complete linkage, as is suggested in [17]. Equation 3 shows the function of the average linkage method between two clusters A and B .

$$H(A, B) = \frac{1}{\|A\|\|B\|} \sum_{n_i \in A, n_j \in B} \text{corr}(n_i, n_j) \quad (3)$$

where the $\text{corr}(n_i, n_j)$ value is calculated with equation 1 and can be taken from the matrix H to simplify the computational processing. $\|A\|$ and $\|B\|$ is the cardinality of the A and B clusters respectively.

Each iteration of the algorithm takes the two clusters with the highest correlation value in the matrix and mixes the objects contained in them to create a new cluster, while storing the state of the different clusters in partition P_0, \dots, P_{N-1} with the aim of knowing the contents of the cluster at any time. In the hierarchical clustering, the final result of the algorithm is a cluster containing all objects. However, in this work each cluster should represent a device at the end of the execution. For this reason, the silhouette coefficient as a measure of validation of clusters was used. The silhouette coefficient measures the similarity index between the elements of a single cluster (cohesion) and the similarity between the elements of a cluster with respect to the others (separation). Unlike Caldelli et al. [17], in our proposal the calculation of the silhouette coefficient is performed for each cluster contained in the P_i partition and not for each pattern noise, as noted in Equation 4.

$$s_j = \max(b_j) - a_j \quad (4)$$

where a_j (cohesion) is the average correlation between all noise patterns within the c_j cluster. b_j (separation) is the average correlation of noise patterns contained in the c_j cluster with respect to noise patterns in the remaining clusters. The nearest neighboring cluster is taken, namely the one with the highest correlation.

For each iteration q of the algorithm a global measure of all the silhouette coefficients calculated from the K clusters is obtained, this is equivalent to averaging the s_j values in q . Equation 5 shows this calculation.

$$SC_q = \frac{1}{K} \sum_{j=1}^K s_j \quad (5)$$

Upon completion of the hierarchical clustering, the SC_q with the lowest value is searched for, which indicates that the

partition P_q^* clusters are at a greater correlation level. The number of clusters at that moment should correspond to the actual number of devices. The aim of storing the partition at each time of the algorithm is to avoid rerunning the clustering because information of all the clusters in each iteration q is known. Next algorithm shows the proposal's pseudocode.

1. Calculate $n^{(i)}$ of each image where $i \in 1, \dots, N$;
2. Generate the similarity matrix $H \in R^{N \times N}$;
3. Foreach $q \in 1, \dots, N - 1$ do
 4. Find cluster $H(k, l)$ with the highest similarity;
 5. Remove the pair of rows and columns corresponding to clusters k and l ;
 6. Calculate the values of the new cluster using average link criteria and add the row and its corresponding column;
 7. Determine the overall silhouette coefficient SC_q ;
 8. Store the partition P_q ;
9. Find the partition where $\min_q(SC_q)$.

IV. EXPERIMENTS AND RESULTS

The experiments were performed with a total set of 1050 photographs from 7 different mobile device camera models. The total set contains 150 photographs from each model. 7 devices are from different manufacturers (Apple iPhone 5, Huawei U8815, LG E400, Samsung GTS5830M, Zopo ZP980, Sony ST25a and Nokia 800 Lumia).

All the images were cropped to 1024×1024 pixels, all images have a horizontal orientation. The scenes of the photographs were chosen randomly, both indoors and outdoors, and they were also taken at different times and places in order to simulate a more realistic scenario. In the extraction of the noise pattern from all images, the zero - mean of rows and columns was used, 3 RGB color channels were converted to a single matrix in grayscale. Additionally, all experiments were conducted using the Wiener filter in the frequency domain.

To measure the degree of certainty in the results, the true positive rate TPR was used. The mean TPR for each of the following experiments is calculated, computing for each cluster the number of photos that have been well classified (TPR of each cluster) and averaging the TPRs of all the resulting clusters (if there are fewer clusters than devices the average takes into account the number of devices). To calculate the TPR of each cluster, the device that has the largest number of images with respect to the total of images by device needs to be identified within the cluster, that being the predominant device cluster, then calculate the percentage of photos that have been well classified for that device in the cluster. Actually, in the vast majority of cases it can be seen that a cluster is associated with one or more devices, as it can be observed in matrices such as the ones in Tables I, II and III. If there are multiple clusters with the same number of photos from a device or a cluster with the same number of photos from several devices and in turn these being the highest, the cluster that is taken as predominant for the device is one chosen among the different options. It may be the case that if there is an extra cluster, a cluster may

not be predominant for any device (see Table II) and its TPR for this cluster is 0. Or there might be one less cluster (see Table III), in this case the association of the cluster to a device will be taken into account and the number of devices will be used to calculate the average, as described above.

In Tables I, II and III there are examples that illustrate the calculation of the TPR for the three cases that may occur.

TABLE I. TPR WITH EQUAL NUMBER OF DEVICES THAN CLUSTERS

Brand - Model	Clusters (%)					Average TPR
	1	2	3	4	5	
Apple Iphone 5	49	0	0	1	0	99.2 %
Huawei U8815	0	50	0	0	0	
LG E400	0	1	49	0	0	
Nokia 800 Lumia	0	0	0	50	0	
Samsung GT5830m	0	0	0	0	50	
TPR by cluster	98	100	98	100	100	

In the results of the experiments 3 possible cases are considered: a) The number of identified clusters is equal to the number of devices, b) the number of identified clusters is higher than the number of devices, and c) the number of identified clusters is lower than the number of devices. Although the first case is ideal, in the second case classifications that do not mix different types of devices in a same cluster can be obtained.

TABLE II. TPR WITH LESS NUMBER OF DEVICES THAN CLUSTERS

Brand - Model	Clusters				Average TPR
	1	2	3	4	
Apple I-iphone 5	100	0	0	0	99 %
Huawei -U8815	0	100	0	0	
LG -E400	0	0	97	3	
TPR by cluster	100	100	97	0	

TABLE III. TPR WITH MORE NUMBER OF DEVICES THAN CLUSTERS

Brand - Model	Clusters (%)				Average TPR
	1	2	3	4	
Apple Iphone 5	100	0	0	0	80 %
Huawei U8815	0	100	0	0	
LG E400	0	0	100	0	
Nokia 800 Lumia	100	0	0	0	
Samsung GT 5830M	0	0	0	100	
TPR by cluster	100	100	100	100	

V. CONCLUSIONS

This paper has made an analysis of the main unsupervised image grouping techniques, which are of utmost importance in digital image forensic analysis. Despite the rise of mobile device cameras these days, there are still few references for unsupervised mobile device image grouping in the state of the art. Most of the works refer to the supervised classification and in many cases they are not focused on mobile device images, which have unique characteristics. The noise added in every photograph by the camera sensor, due to the faults in its

manufacturing process or defects from daily use, has proven to be a reliable source of device identification. Likewise, the calculation of normalized correlation between sensor noises extracted from two or more pictures is also a measure of similarity commonly used in unsupervised image learning techniques, clustering techniques being the ones which obtain the best results. The algorithm of this proposal is based on the combination of a hierarchical clustering and a flat clustering for the separation between clusters. The use of the silhouette coefficient for cluster validation proved to report good results when obtaining high TPRs; also, the number of clusters corresponded to the number of actual devices in most cases. Experiments conducted in this work have revealed a great diversity of situations with regard to the symmetry or not of the photo sets, their size, the number of devices used and the use of devices of the same brand. After all the experiments, it is concluded that the results of the application of the technique are good (92.7% TPR on average for all the experiments).

ACKNOWLEDGMENT

The research leading to these results has been partially funded by the European Union's H2020 Program under the project SELFNET (671672). Part of the computations of this work was performed in EOLO, the HPC of Climate Change of the International Campus of Excellence of Moncloa, funded by MEC and MICINN. This work was supported by the "Programa de Financiación de Grupos de Investigación UCM validados de la Universidad Complutense de Madrid – Banco Santander".

REFERENCES

- [1] J. Hsu, "The Worldwide Mobile Phone Camera Module Market and Taiwan's Industry, 2010 and Beyond", 2010, pp. 1-18.
- [2] R. Baer, "Resolution Limits in Digital Photography: The Looming End of the Pixel Wars - OSA Technical Digest (CD)", in *Proceedings of the Imaging Systems*, Tucson, Arizona United States, June 2010.
- [3] T. Van Lanh, K.S. Chong, S. Emmanuel, M.S. Kankanhalli, "A Survey on Digital Camera Image Forensic Methods", in *Proceedings of the IEEE International Conference on Multimedia and Expo*, Beijing, pp. 16-19, July 2007.
- [4] A. L. Sandoval Orozco, D. M. Arenas González, J. Rosales Corripio, L. J. García Villalba, J. C. Hernandez-Castro, "Techniques for Source Camera Identification", in *Proceedings of the 6th International Conference on Information Technology*, pp. 1-9, May 2013.
- [5] J. Lukas, J. Fridrich, M. Goljan, "Digital Camera Identification from Sensor Pattern Noise", *IEEE Transactions on Information Forensics and Security*, IEEE, 2006, vol. 1 no. 2, pp. 205-214.
- [6] C. F. Eick, N. Zeidat, Z. Zhao, "Supervised Clustering-Algorithms and Benefits", in *Proceedings of the IEEE International Conference on Tools with Artificial Intelligence*, Boca Raton, Florida, USA, pp. 774-776, November 2004.
- [7] J. Fridrich, "Digital Image Forensics", *IEEE Signal Processing Magazine*, IEEE, 2009, vol. 26, no. 2, pp. 26-37.
- [8] C.-T. Li, "Unsupervised Classification of Digital Images Using Enhanced Sensor Pattern Noise", in *Proceedings of the IEEE International Symposium on Circuits and Systems*, Paris, France, May 2010, pp. 3429-3432.
- [9] L. Rokach, "A Survey of Clustering Algorithms", *Data Mining and Knowledge Discovery Handbook*, 2010, pp. 269-298.
- [10] C. T. Zahn, "Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters", *IEEE Transactions on Computers*, IEEE, 1971, vol. C-20, no. 1, pp. 68-86.

- [11] J. D. Banfield, A. E. Raftery, "Model-Based Gaussian and Non-Gaussian Clustering", *Biometrics*, Wiley, 1993, vol. 49, no. 3, pp. 803-821.
- [12] D. H. Fisher, "Knowledge Acquisition Via Incremental Conceptual Clustering", *Machine Learning*, Springer, vol. 2. no. 2, pp. 139-172, 1987.
- [13] J. Vesanto, E. Alhoniemi, "Clustering of the Self-Organizing Map", *IEEE Transactions on Neural Networks*, IEEE, 2000, vol. 11, no. 3, pp. 586-600.
- [14] F. Hoppner, "Fuzzy Cluster Analysis: Methods for Classification", *Data Analysis and Image Recognition, Jossey-Bass Higher and Adult Education Series*, Wiley, 1999.
- [15] S. Z. Selim, K. Alsultan, "A Simulated Annealing Algorithm for the Clustering Problem", *Pattern Recognition*, Elsevier, 1991, vol. 24, no. 10, pp. 1003-1008.
- [16] B.-B. Liu, H.-K. Lee, Y. Hu, C.-H. Choi, "On Classification of Source Cameras: A Graph Based Approach", in *Proceedings of the IEEE International Workshop on Information Forensics and Security*, Seattle, Washington, USA, December 2010, pp. 1-5.
- [17] R. Caldelli, I. Amerini, F. Picchioni, M. Innocenti, "Fast Image Clustering of Unknown Source Images", in *Proceedings of the IEEE International Workshop on Information Forensics and Security*, Seattle, Washington, December 2010, USA, pp. 1-5.
- [18] M. Goljan, J. Fridrich, T. Filler, "Large Scale Test of Sensor Fingerprint Camera Identification", in *Proceedings of the Media Forensics and Security*, vol. 7254, San Jose, California, USA, 2009, pp. 72540I.

Distributed 3D Object Recognition System Using Smartphones

Mustafa Ibrahim¹, Omar El-gendy² and Mohamed Farouk³

Center for Documentation of Cultural and Natural Heritage
Bibliotheca Alexandrina
Giza, Egypt

¹Mostafa_ebrahim87@yahoo.com, ²el-gendy@mcit.gov.eg, ³mfarouk@mcit.gov.eg

Abstract—Object recognition and scene classification are generally considered one of the most important challenges in computer vision community, where, object recognition is a process of finding and identifying objects in a digital image or video sequence. One of the main problems in recognizing 3D object is extracting stable and consistent features vectors under different conditions, such as camera viewpoint, illumination and cluttered background. In addition, Processing and memory capacity of Smartphones still restrict the computational capacity of object recognition programs. In this paper, we propose a distributed 3D object recognition system to overcome computational capacity problem and improve scalability of objects that will simply be recognizable. The paper also proposes the use of k-Nearest Neighbors classifier with Speeded Up Robust Features algorithm to solve the problem of extracting stable and consistent features vectors. The system is remarkably capable of adapting to different network configurations and the wireless bandwidth, and improving the performance of recognizing multiple 3D objects using Smartphones devices.

Keywords—Scale Invariant Feature Transform; Speeded Up Robust Features; k-Nearest Neighbors.

I. INTRODUCTION

One of the most significant developments in the last decade is the applications of 3D object recognition. Object recognition is a computer technology related to computer vision and image processing that deals with finding and identifying instances of semantic objects of a certain class in digital images or video sequence. The main factors that affect the accuracy of 3D object recognition systems are the variability in the illumination and the pose of the objects, in addition to time delay. The presence of these factors in recognizing 3D objects can lead to diminishing recognition reliability [1].

Obtaining 2D images from 3D scenes is the reverse process of reconstruction 3D Model from multiple 2D images. Thus, it can handle any 3D object as a sequence of 2D images.

Selecting the most consistent, stable and reliable feature extraction technique and appropriate classifier for classifying number of categories, which contain large number of features is a hard task that will be overcome by using Speeded Up Robust Features (SURF) algorithm in extracting features and K-Nearest Neighbor classifier in classifying objects.

Smartphones devices are very limited in executing high computational capacity programs such as objects recognition and image classifications programs, that's because of the high computational capacity required for this type of applications,

in addition to low capacity of memory and processors of smartphones devices up to this day.

Offloading computations from smartphones to remote cloud resources has recently been rediscovered as a technique to enhance the performance of smartphone applications, while reducing the energy usage [2].

The rest of this paper is structured as follows: the next section presents the problem statement. Section 3 talks about some related works. Section 4 and 5 discuss the framework of the proposed recognition system. Section 6 explains the proposed distributed system. The comparative experiments and results are discussed in Section 7. Finally, the paper will be concluded in Section 8.

II. PROBLEM STATEMENT

Extracting stable and consistent features under different condition such as reflections, illumination and camera view point has been considered one of the main challenges in recognizing 3D objects. Besides that, Smartphones still restrict the computational capacity of object recognition programs, while, 3D object recognition programs involve complex mathematical calculations and they require powerful memory and processor to cover their computational needs. Therefore, in this paper the proposed distributed 3D object recognition system can handle the computational capacity and scalability problems of smartphone resources. Moreover, using robust

and fast algorithm such as SURF can solve the recognition problems with high accuracy.

III. RELATED WORK

The most common way to tackle 3D detection is to represent a 3D object by a collection of independent 2D appearance models [3, 4, 5, 6, 7], one for each viewpoint. Several authors augmented the multi-view representation with weak 3D information by linking the features or parts across views [8, 9, 10, 11, 12]. This allows for a dense representation of the viewing sphere by morphing related near-by views [13], since these methods usually require a significant amount of training data.

Two general approaches have been taken to solve 3D recognition problem: pattern recognition approaches and feature-based geometric approaches. The first approach uses low-level image appearance information to locate an object, while the second constructs a model for the object to be recognized, and matches the model against the photograph.

The groundbreaking work of Schmid and Mohr showed that invariant local feature matching could be extended to general image recognition problems in which a feature was matched against a large database of images. They also used Harris corners to select interest points, rather than matching with a correlation window. The Harris corner detector is very sensitive to changes in image scale, so it does not provide a good basis for matching images of different sizes [14].

David G. Lowe extended the local feature approach to achieve scale invariance using Scale Invariant Feature Transform (SIFT) algorithm. This work also described a new local descriptor that provided more distinctive features while being less sensitive to local image distortions such as 3D viewpoint change. The high dimensionality of Lowe descriptors was a drawback of SIFT algorithm [14].

Speeded Up Robust Features (SURF) algorithm, on the other hand, is designed for much faster scale-space extraction. The detection of extrema is located on the determinant of Hessian approximated by Haar-wavelets. The descriptor is based on the polarity of the intensity changes. Sums of the gradient (oriented with the main orientation of the keypoints) and the absolute of gradient in horizontal and in vertical direction are computed [15].

However, some computation intensive applications cannot be run on smartphones since their computing power and battery life are still limited for such resources. Some of these applications including video encoding/decoding, image recognition, and 3D graphics rendering, could take a significant amount of time due to their computationally intensive nature. Processors on mobile devices are gradually getting faster year by year; however, without aid from special

purpose hardware, they may not be fast enough for those computationally intensive applications [16].

Recently, it has been rediscovered that offloading computation using the available communication channels to remote cloud resources can help to reduce the pressure on the energy usage. Furthermore, offloading computation can result in significant speedups of the computation, since remote resources have much more compute power than smartphones [2].

IV. SPEED UP RUBOST FEATURES ALGORITHM

Feature extraction is one of the most important steps in image pattern recognition tasks. As happens with any pattern recognition algorithm, the performance of recognition algorithm strongly depends on the feature extraction method and the classification systems used to carry out recognition tasks [17].

Scale Invariant Feature Transform (SIFT) is an approach for detecting and extracting local features descriptors that are reasonably invariant to changes in rotation, scaling, lighting conditions and small changes in view point. SIFT features are also very resilient to the effects of "noise" in the image [17]. Generally, the high dimensionality of the descriptor is a drawback of SIFT algorithm at the matching step. For on-line applications relying only on a regular PC, each one of the three steps (detection, description, matching) has to be fast [18].

Speeded Up Robust Features (SURF) is a robust local feature detector, first presented by Herbert Bay et al. in 2006, that can be used in computer vision tasks like object recognition or 3D reconstruction. It is partly inspired by the SIFT descriptor. The standard version of SURF is several times faster than SIFT and claimed by its authors to be more robust against different image transformations than SIFT [17].

In SURF algorithm, the most valuable property of an interest point "Detector" is its repeatability. The repeatability expresses the reliability of a detector for finding the same physical interest points under different viewing conditions. Next, the neighborhood "Descriptor" of every interest point is represented by a feature vector. This descriptor has to be distinctive and at the same time robust to noise. The dimension of the descriptor has a direct impact on the time this takes, where, less dimensions are desirable for fast interest point matching [18].

In SURF feature vector. In order to bring in information about the polarity of the intensity changes, we also extract the sum of the absolute values of the responses, $|d_x|$ and $|d_y|$. Hence, each sub-region has a four-dimensional descriptor vector v for its underlying intensity structure

$$v = (\sum d_x; \sum d_y; \sum |d_x|; \sum |d_y|). \quad (1)$$

Where d_x is the Haar wavelet response in horizontal direction, and d_y is the Haar wavelet response in vertical direction.

Concatenating this for all 4 x 4 sub-regions, this results in a descriptor vector of length 64. Figure 1 shows the properties of the descriptor for three distinctively different image-intensity patterns within a sub-region [18].

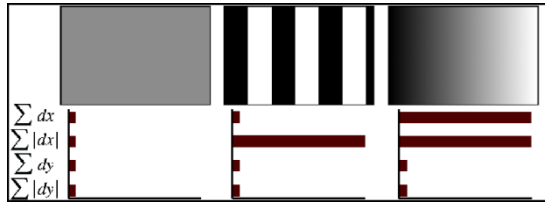


Fig. 1. The descriptor entries of a sub-region represent the nature of the underlying intensity pattern. Left: In case of a homogeneous region, all values are relatively low. Middle: In presence of frequencies in x direction, the value of $\sum |d_x|$ is high, but all others remain low. If the intensity is gradually increasing in x direction, both values $\sum d_x$ and $\sum |d_x|$ are high [18].

Table (1) and (2) show that, SIFT has detected more number of features compared to SURF but it is suffered with speed. SIFT is slow and not good at illumination changes, while it is invariant to rotation, and scale changes. SURF is fast and has good performance as much as SIFT.

TABLE 1. COMPARISONS OF RESULTS OF SIFT AND SURF ALGORITHM [19]

Algorithm	Detected Feature Points		Matching feature point	Feature matching Time
	Image1	Image2		
SIFT	892	934	41	1.543 s
SURF	281	245	28	0.546 s

TABLE 2. COMPARISONS OF RESULTS OF SIFT, PCA-SIFT AND SURF ALGORITHM [20]

Algorithm	Time	Scale	Rotation	Blur	Illumination
SIFT	common	best	best	best	common
PCA-SIFT	good	common	good	common	good
SURF	best	good	common	good	best

V. K-NEAREST NEIGHBOR CLASSIFIER

K-Nearest Neighbor (KNN) is one of the most popular algorithms for pattern recognition, which has been proven to be a simple and powerful recognition algorithm. Many researchers have found that the KNN algorithm accomplishes very good performance in their experiments on different data sets [21].

K-Nearest Neighbor (KNN) is a supervised learning algorithm and it is a non-parametric method for classifying objects based on closest training examples in the feature space. In statistics, the term non-parametric covers techniques that do not rely on data belonging to any particular distribution [17].

The KNN classification algorithm predicts the test sample's category according to the K training samples which are the nearest neighbors to the test sample, and then judges it to that category which has the largest category probability [21]. The process of KNN algorithm to classify sample X is as follows [21]:

- Suppose there are j training categories C_1, C_2, \dots, C_j and the sum of the training samples is N after feature reduction, they become m -dimension feature vector.
- Make sample X to be the same feature vector of the form (X_1, X_2, \dots, X_m) , as all training samples.
- Calculate the similarities between all training samples and X . Taking the i^{th} sample $d_i (d_{i1}, d_{i2}, \dots, d_{im})$ as an example, the similarity $\text{SIM}(X, d_i)$ is as follows:

$$\text{SIM}(X, d_i) = \frac{\sum_{j=1}^m X_j \cdot d_{ij}}{\sqrt{\left(\sum_{j=1}^m X_j\right)^2} \cdot \sqrt{\left(\sum_{j=1}^m d_{ij}\right)^2}} \quad (2)$$

- Choose k samples which are larger from N similarities of $\text{SIM}(X, d_i)$, ($i=1, 2, \dots, N$), and treat them as a KNN collection of X . Then, calculate the probability of X that belongs to each category respectively with the following formula.

$$P(X, C_j) = \sum_d \text{SIM}(X, d_i) \cdot y(d_i, C_j) \quad (3)$$

Where $y(d_i, C_j)$ is a category attribute function, which satisfied

$$y(d_i, C_j) = \begin{cases} 1, & d_i \in C_j \\ 0, & d_i \notin C_j \end{cases}$$

Judge sample X to be the category which has the largest $P(X, C_j)$ as shown in figure 2.

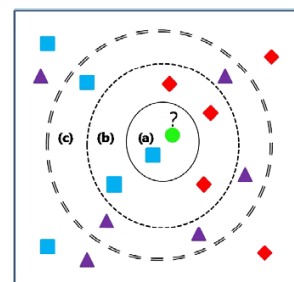


Fig. 2. K-NN Classification. At the query point of the circle depending on the k value of 1, 5, or 10, the query point can be a rectangle at (a), a diamond at (b), and a triangle at (c)[22].

The accuracy of the k-NN algorithm can be severely degraded by the presence of noisy and irrelevant features, or if the feature scales are not consistent with their importance. So, extracting consistent and relevant features using SURF algorithm can help k-NN algorithm in classifying 3D objects with high accuracy.

VI. DISTRIBUTED 3D OBJECT RECOGNITION SYSTEM

This paper follows another line of research on building distributed 3D object recognition system. This distributed system is a software system in which, components located on networked computers communicate and coordinate their actions by passing messages [23]. Therefore, the proposed system tries to employ the power of using distributed system in recognizing objects.

The proposed system for recognizing 3D objects consists of three steps. The first step is a sampling procedure that captures a finite set of candidate 3D locations in order to avoid the high computational cost of considering every potential location. The second step is extracting the most stable and consistent features of each captured image and merging them to create the marker of a captured object. The last step is matching the new image that was captured by a smartphone with stored markers list in a workstation and returns the result back to the smartphone.

Figure 3 illustrates a scenario for capturing multi scene of the same object from different viewpoints. For each 3D object, 18 (2*9) different vantage points have been selected to measure the 3D appearance of this object.

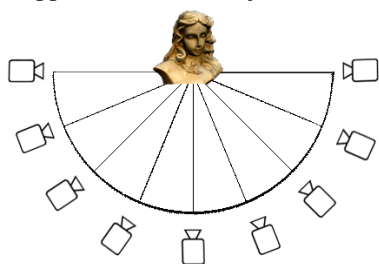


Fig. 3. Capture multi scenes of the same object from different multiple viewpoints

After extracting features of each object using SURF algorithm (local feature vectors for each image) and merging them, the obtained marker of each object has been ready to use in the matching process. Allocate all markers that have been created from extracting and merging features processes on a workstation to be used later in the features matching process.

Figure 4 shows the main architecture of the proposed distributed system. It consists of a workstation, wireless router and many smartphones. Although this is very simple distributed system architecture, it is a very effective system.



Fig. 4. The architecture of distributed 3D objects recognition system using smartphones.

The workstation will be connected to the router using Ethernet cable “Wired network”, actually, wireless network can be used to connect the workstation and router instead of wired network, but it is better to connect them using a wired network because the wired network is faster and reliable than the wireless network.

In the proposed distributed system, the workstation can be configured as follows, install one of the editions of Windows Vista or Windows 7 on which IIS 7 (Internet Information Services) and above is supported before you proceed. Also be sure that you have administrative user rights on the computer.

After configuring the workstation, web services technologies can be used as a method of communication among many different electronic devices (smart phones and workstation) over a network.

Web services provide an infrastructure for maintaining a richer and more structured form of interoperability between clients and servers. In particular, web services allow complex applications to be developed by providing services that integrate several other services [23].

In The proposed distributed system, the developed web service is responsible for executing three procedures. The first procedure is loading and caching all markers of all objects that have been allocated on workstation web service (once any smartphone connects to the web service). The second procedure is receiving images that have been captured using smart phones and extracting their features using SURF algorithm. The last procedure is matching the extracted features with the loaded

markers feature using KNN algorithm and sending result back to the smart phone as shown in figure 5.

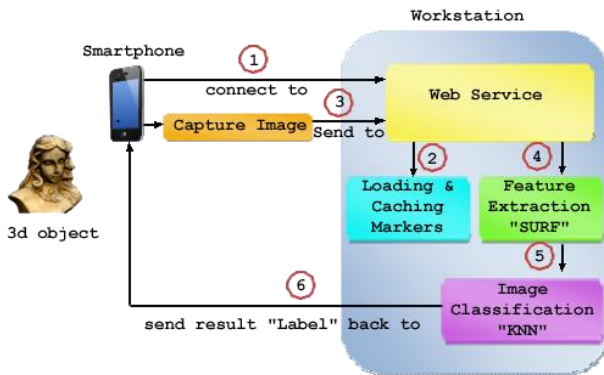


Fig. 5. Data flow diagram of the proposed distributed 3D object recognition system using smartphones.

VII. EXPERIMENTS AND RESULTS

In the experiments of the proposed system, we use some smartphone devices in testing system such as:

- Lenovo Tablet A3000 (OS: Android, v4.1, Processor: Quad-core 1.2 GHz Cortex-A7, RAM: 1 GB and Camera: 5 MP, 2592 x 1936 pixels, autofocus).
- HTC mobile phone Desire 816 (OS: Android, v4.4.2, Processor: Quad-core 1.6 GHz Cortex-A7, RAM: 1.5 GB and Camera: 13 MP, 4160 x 3120 pixels).
- Samsung mobile phone Galaxy Ace 3 (OS: Android, v4.2, Processor: Dual-core 1 GHz Cortex-A9, RAM: 1 GB and Camera: 5 MP, 2592 x 1944 pixels, autofocus).
- Samsung mobile phone Galaxy A5 Duos (OS: Android, v4.4.4, Processor: Quad-core 1.2 GHz Cortex-A53, RAM: 2 GB and Camera: 13 MP, 4128 x 3096 pixels, autofocus).

The configurations of the workstation which is responsible for executing all features extracting and matching procedures is HP Pro 3300 powered by the 2nd generation Intel® Core™ i5 processors running at 2.5 GHz, RAM: 4 GB, OS: Windows 7 professional and IIS 7 server. Both the smartphone devices and the workstation are connected within the same network segment via Wi-Fi 802.11g. D-LINK DSL-2640T router supports wireless speed up to 54 Mbps and interoperability with 802.11b wireless devices on the 2.4GHz frequency band.

The proposed system has been evaluated by real data provided by the Center for Documentation of Cultural and Natural Heritage (CULTNAT) as shown in figure 6. The experiment tested 440 different images of 11 objects whose sizes range from 120 KB to 200 KB and their dimensions are 352 x 288 pixel. Those images are captured from different distances and multiple view points.

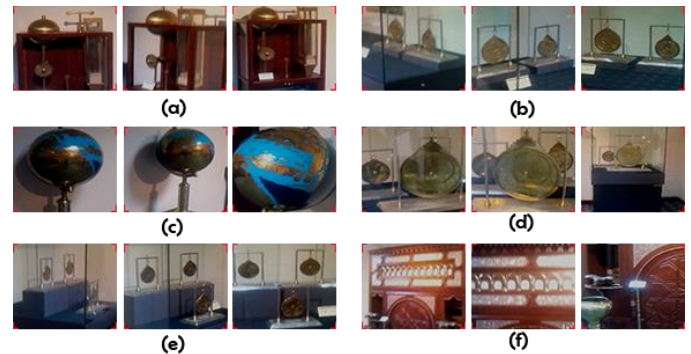


Fig. 6. Examples of different 3D objects have been captured from different distances and multiple viewpoints. In (a), Sand Clock model, In (b), The astrolabe of As-Sahli model, In (c) world map of the geography of the caliph ma'mun, In (d), The newest astrolabe model, In (e) The astrolabe of gafar al muktafi model, and In (f), Water clock Model

The version of OpenCV is 2.2.0 which is used in the workstation side. The main classes have been used are SURF Detector class for extracting features and Flann class for matching features.

During the experiment, no user applications on the smartphone devices other than the proposed system are launched.

Using the proposed system we conducted several experiments and here we represent some results of our experiments. Figure 7 and 8 appear the results of recognizing 11 categories with different K values using K-Nearest Neighbor algorithm and 220 images as testing data. In this experiment we compare the results of recognizing different objects when k equals 2, 4, 8, 16, 32, 64 and 128. For readability we rename the used objects as class 0, class 1 and so on, in addition to we separate the results in two charts. Figure 9 represents the average of success of this experiment. The best result was obtained when k = 64.

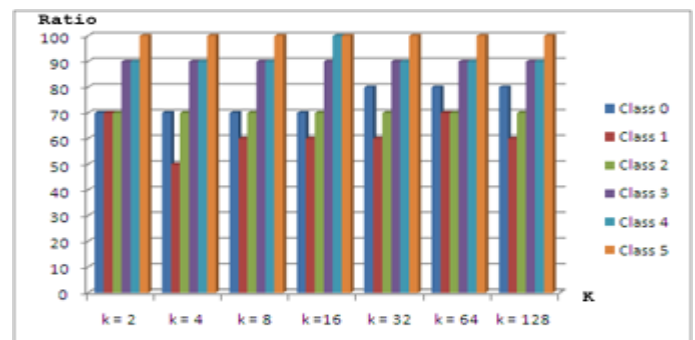


Fig. 7. Chart of recognizing the first 6 classes using different k values
 (Where K=2, 4, 8, 16, 32, 64,128)

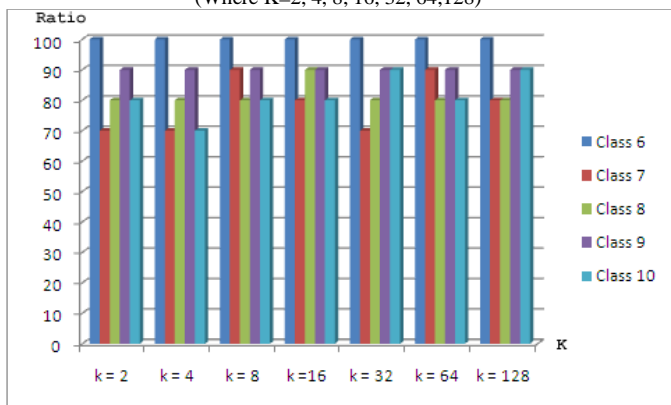


Fig. 8. Chart of recognizing the second 5 classes using different k values
 (Where K=2, 4, 8, 16, 32, 64,128)

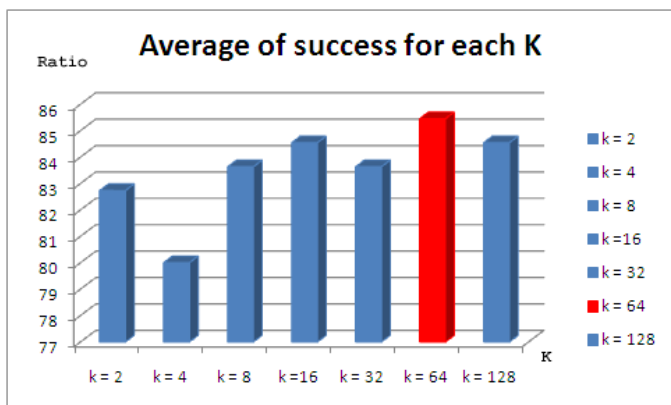


Fig. 9. The average of success of recognizing 11 objects using different k values
 (Where K=2, 4, 8, 16, 32, 64,128)

The previous experiments lead us to conduct other experiments with different values of k to help us in deciding which k must be used. Figure 10 and 11 represent the results of testing data when k=50, 100, 150, 200, 250 and 300. Figure 12 represents the average of success of this experiment. The best result was obtained when k =150. Increasing the value of k does not improve recognition performance but this consume more time than other experiments, so k=150 is the best one.

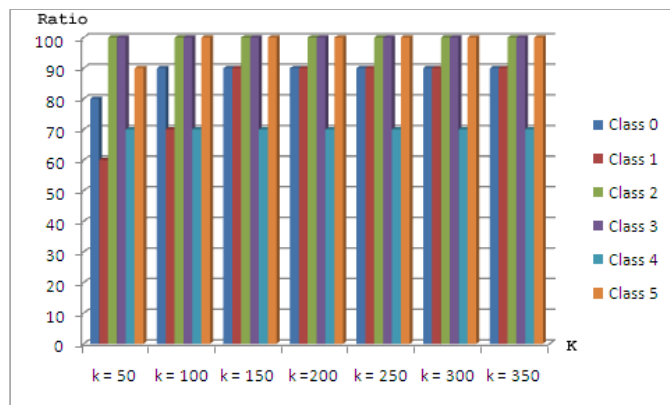


Fig. 10. Chart of recognizing the first 6 classes using different k values
 (Where K=50, 100, 150, 200, 250, 300,350)

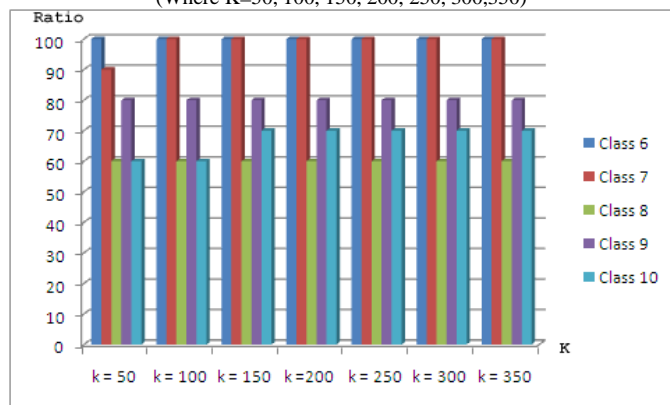


Fig. 11. Chart of recognizing the second 5 classes using different k values
 (Where K=50, 100, 150, 200, 250, 300,350)

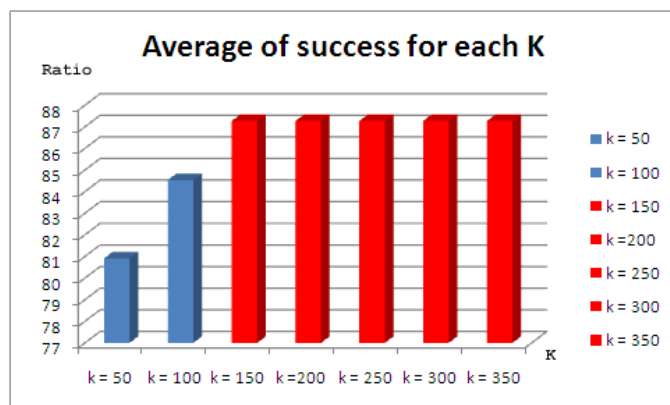


Fig. 12. The average of success of recognizing 11 objects using different k values (Where K=50, 100, 150, 200, 250, 300,350)

VIII. CONCLUSION

In this paper we proposed a distributed 3D object recognition system using smartphones. The system uses selective Speeded Up Robust Features algorithm to extract salient properties of appearance descriptors of local image patches. Furthermore, K-Nearest Neighbor classifier has been used as a simple and powerful recognition algorithm.

Smartphones are only responsible for capturing images of 3D objects and sending them to the distributed workstation through a wireless network. All other processes of extracting and matching features are performed by the distributed workstation. Consequently, the proposed distributed system can handle computational capacity problem of smart phones and improve scalability of objects that will accurately be recognizable. 440 images have been used as a simple sample of testing data. Our experiments on a variety of 3D objects demonstrated the effectiveness of the proposed system.

IX. FUTURE WORK

A recent work shows that, extracting feature vectors can be accelerated using FREAK descriptors [24]. It will be interesting to try such methods to make our approach faster. Moreover, the accuracy could be improved by using support-vector-machines (SVM) classifier instead of K-nearest-neighbor classifier, as suggested by another recent work [22].

REFERENCES

- [1] Krizaj Janez, Štruc Vitomir, Dobrisek, and Simon. "Robust 3D Face Recognition", journal of Electrical Engineering and Computer Science (Elektrotehniški Vestnik), Ljubljana, Slovenia, 2012.
- [2] Roelof Kemp, Nicholas Palmer, Thilo Kielmann and Henri Bal, "A computation offloading framework for smartphones", Conference on Mobile Computing, Applications, and Services, Springer Berlin, Heidelberg, Germany, 2012.
- [3] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models". IEEE Transactions on Pattern Analysis and Machine Intelligence, Chicago, 2010, 32, pp. 1627–1645.
- [4] H. Schneiderman and T. Kanade, "A statistical method for 3d object detection applied to faces and Cars", Conference on Computer Vision and Pattern Recognition, San Francisco, 2000, pp. 1746–1759.
- [5] A. Torralba, K. P. Murphy, and W. T. Freeman, "Sharing visual features for multi-class and multi-view object detection". IEEE Transactions on Pattern Analysis and Machine Intelligence, Chicago, 2007, 29, pp. 854–869.
- [6] C. Gu, and X. Ren, "Discriminative mixture-of-templates for viewpoint classification". European Conference on Computer Vision, Crete, Greece, 2010, pp. 408–421.
- [7] B. Pepik, M. Stark, P. Gehler, and B. Schiele, "Teaching 3d geometry to deformable part models", Computer Vision and Pattern Recognition, IEEE Conference, 2012.
- [8] A. Kushal, C. Schmid, and J. Ponce, "Flexible object models for category-level 3d object recognition", Conference on Computer Vision and Pattern Recognition, San Francisco, 2007.
- [9] A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, B. Schiele, and L. V. Gool, "Toward multi-view object class detection", Conference on Computer Vision and Pattern Recognition, San Francisco, 2006.
- [10] D. Hoiem, C. Rother, and J. Winn, "3D layout for multi-view object class recognition and Segmentation", Conference on Computer Vision and Pattern Recognition, San Francisco, 2007.
- [11] Sun, M., Su, H., Savarese, S., and Fei-Fei, L., "A multi-view probabilistic model for 3d object Classes", Conference on Computer Vision and Pattern Recognition, San Francisco, 2009.
- [12] N. Payet, and S. Todorovic, "Probabilistic pose recovery using learned hierarchical object models". International Conference on Computer Vision Barcelona, Spain, 2011.
- [13] H. Su, M. Sun, L. Fei-Fei, and S. Savarese, "Learning a dense multi-view representation for detection, viewpoint classification and synthesis of object categories", International Conference on Computer Vision, Kyoto, Japan, 2009.
- [14] David G. Lowe, "Distinctive image features from scale-invariant keypoints", International journal of computer vision, 2004, pp. 91–110.
- [15] D. Marimon, A. Bonnini, T. Adamek, and R. Gimeno, "DARTs: Efficient scale-space extraction of DAISY keypoints", Conference on Computer Vision and Pattern Recognition, San Francisco, 2010.
- [16] Imai and Shigeru, "Task offloading between smartphones and distributed computational resources", Diss. Rensselaer Polytechnic Institute, 2012.
- [17] M. Labib, M. Fakhr, and M. Ali, "Large scale linear coding for image classification", 1st ed, Germany, Deutschland, 2014.
- [18] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool, "SURF: Speeded up robust features", Computer vision and image understanding journal, 2008, vol. 110.
- [19] P. M. Panchal¹, S. R. Panchal², and S. K. Shah, "A Comparison of SIFT and SURF", International Journal of Innovative Research in Computer and Communication Engineering, Tamilnadu, India, 2013, vol. 1, Issue 2.
- [20] Luo Juan and Oubong Gwon, "A Comparison of SIFT, PCA-SIFT and SURF", International Journal of Image Processing, Malaysia, 2009, vol. 3, Issue 4.
- [21] N. Suguna, and Dr. K. Thanushkodi, "An Improved k-Nearest Neighbor Classification Using Genetic Algorithm", International Journal of Computer Science Issues, Mahebourg, 2010, Vol. 7, Issue 4, No 2.
- [22] Kim, Junho, Byung-Soo Kim, and Silvio Savarese. "Comparing image classification methods: K-nearest-neighbor and support-vector-machines". Proceedings of the 6th WSEAS international conference on Computer Engineering and Applications, 2012.
- [23] George Coulouris, Jean Dollimore, Tim Kindberg, and Gordon Blair, Distributed Systems Concepts and Design, 5th ed. Boston, Addison-Wesley, 2011.
- [24] J. Krizaj, V. Struc, S. Dobrisek, D. Marcetić, and Ribarić, "SIFT vs. FREAK: Assessing the usefulness of two keypoint descriptors for 3D face verification", IEEE, Information and Communication Technology, Electronics and Microelectronics, 2014.

A Survey on Digital Image Steganography

Zaid Al-Omari

Department of Computer Science
Yarmouk University
Irbid, Jordan
Zaidcs2008@gmail.com

Ahmad T. Al-Taani

Department of Computer Science
Yarmouk University
Irbid, Jordan
ahmadta@yu.edu.jo

Abstract— The fast growth in communication technologies and the increased availability of the public networks (Internet) facilitated data transfer. However the public communication channels are vulnerable to security attacks that may lead to unauthorized access to some information. Encryption has been used to persist and prevent these attacks. But when the information is decrypted it will be exposed to the attackers again and it will not have any security protection. Steganography is the science of embedding the secret messages inside other medium files (text, audio, image, and video) in a way that hides the existence of the secret message at all. Steganography applies an embedding process in which the redundant bits of the medium are replaced by the bits of the secret message. Image Steganography is the field of steganography in which the medium that used to carry the secret data is a digital image. Image Steganography is an important area of research in the recent years. This article reviews the steganography based on digital images; illustration of the concept and the common approaches are discussed. Also Steganalysis which is the science of attacking steganography will be briefly discussed.

Keywords— *digital image steganography; spatial domain embedding; transform domain embedding; cover-image; stego-image; wavelet transform; Steganalysis*

I. INTRODUCTION

Steganography comes from the Greek words Steganos (Covered) and Craptos (Writing) [16], Steganography has been used thousands of years ago, tattoos or invisible ink are an examples of the old techniques for steganography; in the 5th century BC a Roman general shaved a slave's head and tattooed a message on it and after his hair grew back he sent him to deliver the hidden message (tattooed on his hair) [17].

The concept of "What You See Is What You Get" with respect to digital images is no longer accurate. Images may be more than what we can see using our Human Visual System (HVS); because it can hold an embedded data (secret messages) that cannot be seen if a steganography algorithm was applied to it.

Nowadays Steganography can be defined as the science of communicating a secret data by embedding it in a multimedia carrier such as image, audio or video files to produce a stego-file that contains the secret data and attempting to foil the human visual system (HVS) and the steganalysis algorithms.

In steganography there exist two types of materials the first one is the message which is the secret data that will be embedded and transferred on the second material (carrier) which is the material that will hold the message. Modern steganography tries to be detectable only if secret information is known; secret keys [19] are shared between the sender and receiver.

Cryptography is another method that is used for secret communication. Cryptography involves converting a message text into an unreadable cipher. Cryptography differs from Steganography in that steganography hides the message so it cannot be seen while the cryptography techniques scramble the message so it cannot be understood. However both of them can be combined to produce better security and protection of the message [20]. In this approach if the steganography fails and the message was detected, the steganalyser gets a message which was encrypted using cryptography techniques so that the message cannot be understood unless the cryptography technique used was also detected.

Many surveys on image steganography have been published, the most popular one [38] was published six years ago; it provides a comprehensive view of image steganography

and includes the literature that have been published until the time the paper was published. But now it may be considered out of date because there are too many contributions published from that time, these new publications needs to be included into a new survey. Also some common image steganography techniques in spatial and transform domains have been discussed by other surveys [21, 22] but both of them concentrated on defining the steganography and steganalysis types and provided a classifications of steganography and steganalysis techniques, also they did not provide neither a comparison between the related publication nor the advantages and disadvantages for each research that are discussed. Other surveys [39, 40] briefly discussed the image steganography's definition, domains and techniques in a very concise form without discussing the huge amount of contributions of researchers on this field. In this paper we investigated the most recent researches on image steganography, critically analyzed them, provided comparisons if applicable and related the publications that have been emerged from previous same publication.

The goal of this paper is to critically analyze the various image steganography techniques that have been proposed in the recent few years and also providing an overview of the digital image steganography major domains. By studying the literature on this topic especially the recent published researches, we can start our research to produce a new technique or provide an enhancements to existing one.

In this paper, also we review steganography research based on digital images. The rest of the paper is organized as follows: Section II discusses the digital image steganography. Section III defines the major domains of image steganography; the spatial domain and transform domain embedding methods, also the hybrid approaches are discussed briefly. In section IV, the performance specification of image steganography is discussed. Section V provides critical analysis of the main discussed literature. Finally, Conclusions are drawn in section IV.

II. IMAGE STEGANOGRAPHY

The rapid development and enhancement on imaging technologies and devices that resulted on producing a high resolution and accurate images, also the fast growth and spread of the Internet has paved the way to transfer huge amounts of information including such images, many of these information needs to be transmitted in a secure and private way. Therefore, steganography gets a role on the stage of information security (information hiding).

In digital image steganography almost all file types can be used but images have been proven to be the most suitable for embedding because of their high degree of redundancy [18].

When we study, discuss or develop image steganography systems there are three factors [17] must be considered:

- Capacity: refers to the amount of data that can be embedded into the cover image, sometimes it is called the payload.

- Robustness: its resistance to various image processing and compression.
- Security, imperceptibility or undetectability: minimizing the modifications to the cover image so that the resulted stego image can resist the steganalysis and the HVS.

Fig. 1, Illustrates an overview of an image steganography system. The message is embedded into the cover image by using stego image encoder and a shared key to produce the stego image that will be transferred to the receiver and then the system stego decoder will use the same key to extract the secret message.

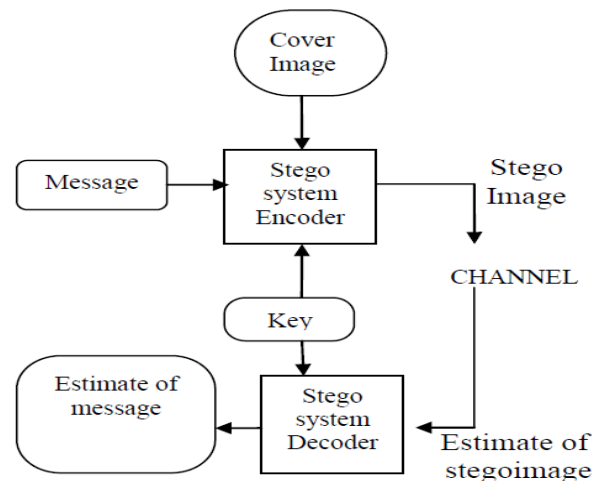


Fig. 1. Overview of Steganographic system.[39]

III. IMAGE STEGANOGRAPHY DOMAINS

Image steganography has huge and continuous research contributions from the researchers around the world. As a result to these contributions from the research community; large number of embedding techniques have been proposed, some of them have been applied to gray-scale images and others to colored images. All of these techniques modify the cover image in different ways to embed the secret message bits. The common goals of all techniques are making embedding rate as the high as possible while preserving the undetectability against the steganalysis attacks as much as possible.

All the proposed techniques can be classified into three major domains: spatial domain embedding, transform domain embedding or hybridization between them or with other approaches.

A. Spatial Domain Embedding

Spatial domain steganography is based on physical location of pixels in an image. It involves encoding at the level of LSBs. In this technique only the least significant bits of the cover object is replaced without modifying the complete cover object. It is the simplest method for data embedding but is weak in resisting attacks such as compression and transforms [21]. However its payload is large.

Least Significant bit Replacement method (LSB): LSB is the most common approach, it is a simple approach in which some or all of the LSBs of the image pixels are changed to a bit of the secret message. [1] used LSB in colored images by replacing only the blue color bits to provides more efficiency and less distortion to the cover image. Fig. 2 provides an example of the LSB embedding method.

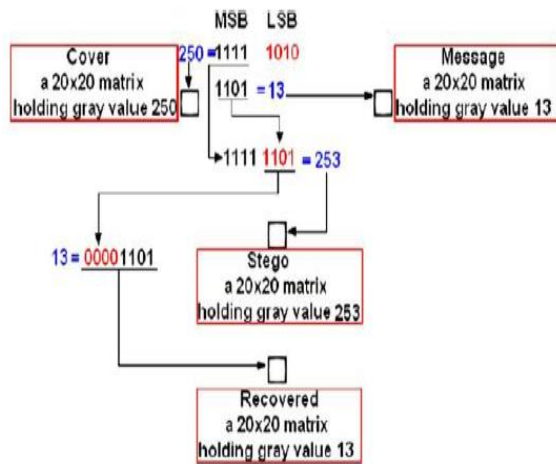


Fig. 2. LSB embedding method.[38]

References [2-5] take the advantage of edge detection techniques to increase the capacity (payload) of the secret message that the cover image can carry without affecting undetectability (imperceptibility).

Reference [2] proposed a novel steganography scheme based on LSB and a hybrid edge detector by combining a fuzzy edge detector and canny edge detector to select various numbers of LSB's for each cover image pixel that can be used to be replaced with the secret message bits, the results shows that it achieves high embedding payload and preserving the stego image quality. While [4] takes advantage of sharp areas in images to hide large amount of information inside it, it also used a hybrid edge detector as [2] but the two detectors used are Sobel and The Laplacian filters. The results shows that the method proposed by [4] increases embedding capacity compared to [2] while the image quality is almost the same.

Least Significant bit Matching algorithm (LSBM): In LSB replacement which embeds secret data by replacing the LSB's bits of the cover image with the secret message bits directly which leads to imbalance in the embedding because modifies even pixel values only and leaves odd values unchanged it can be detected easily by current methods [6-10]. However LSB matching also modifies the LSBs of the cover image but it not simply replace them as LSB replacement do, it checks if the message bit does not match the cover image LSB bit it randomly increase or decrease one from the value of the cover pixel which increase the security.

Least Significant bit Matching Revisited algorithm (LSBMR): Proposed by [11] as an enhancement to LSBM algorithm to minimize the changes that should be made on the cover image while still embedding the same payload by

applying a new method in which a choice to set a binary function of two cover image pixels to the desired value; the experimental results shows that it eliminates the imbalance that results with LSB replacement and also decreases the modifications per pixel compared to the LSB matching.

References [12-15] have used the LSBMR algorithm proposed by [11] by combining it with adaptive schemes and compared their work to each other. based on analysis and extensive experiments [12] noticed that the LSBMR does not take into consideration the relationship between the cover image and the size of the secret message to be embedded which will cause the smooth/flat to become distorted after hiding the message even at low embedding rates that implies poor visual quality and low security specially if the cover image contains many smooth regions; he proposed a novel scheme which can embed the secret message into the sharper edge regions adaptively based on a threshold value that depends on the size of the message and the gradients of the content edges by using an edge adaptive scheme to select the embedding regions based on the size of the secret message and the differences between consecutive pixels in the cover image. The experimental results prove that this scheme preserves the statistical and visual features of the stego image.

References [13, 14] and [15] discovers that on [12] the pulse distortion introduced in the histogram of absolute difference of pixel pairs can be detected [13] have performed extensive experiments on the scheme proposed at [12] and finds that the Discrete Fourier Transform (DFT) spectrum of pixel pairs differences histogram still reveals the existence of secret data even at a low embedding rates.

Reference [13] proposed a block-based adaptive steganography combined with the LSBMR in which the cover image with size $m \times n$ and 8 bits/pixel is divided into non overlapping $B \times B$ blocks where B is a random number, and the threshold value was determined by the size of the secret message and the total quantity of blocks that the total pixel-pairs' absolute difference. The experimental results proves that it is more secure and keeps the visual quality better than EA-LSBMR in higher embedding rates and it outperforms the other stenographic methods including LSBM, LSBMR, PVD and IPVD.

To solve the pulse distortion to the histogram of the absolute difference of the pixel pairs resulted from [12] [14] proposed a steganalytic method based on B-Spline fitting to detect stego images with low embedding rate like the EA-LSBMR. Experimental results shows that the new steganalytic method provides better performance than the other state of art steganalyzers also it can estimate the threshold that was used in the embedding process accurately. While [15] benefits from [14] to produce an improved algorithm for EA-LSBMR in which the adjacent pixel pairs for data hiding are selected randomly to solve the pulse distortion weakness discussed on [14] regarding the work presented on [12]. While [15] divided the image into 3×3 non-overlapping blocks (similar to work presented on [13] but with fixed size rather than random way) and the pixel pairs are randomly selected from each of these blocks, this technique cannot be attached because the pixel

pairs from each block are selected randomly also the introduced pulse distortion cannot be discovered any more as the experimental results shows. Thus the Improved EA-LSBMR algorithm resists the steganalyzer proposed by [14] which was designed to detect pulse distortion by using a new scheme for selecting adjacent pixel pairs.

Multiple Bit-planes Based Steganography (MBP): Proposed by [23] in which they developed the bit-plane complexity segmentation (BPCS) steganography. This method was designed to be secure against classical steganalysis attacks. But it is limited to compressed images only.

Modulo Operation Based Steganography: it is an adaptive steganography scheme proposed by [24] which introduces a Multiple-Based Notational System (MBNS) depending on Human Vision Sensitivity (HVS). In this system the hiding capacity for each image pixel takes into consideration the factor of human visual sensitivity. This system allows more secret data to be hidden and the produced stego image degradation is very invisible to human eye.

A novel approach proposed by [25]: a spatial domain steganography method for gray-scale images. In this approach the cover image is divided into equally size blocks and the hidden message bits was embedded in the edge of the blocks. The embedding process depends on the number of ones in left four bits of the pixel. This method was compared to PVD and GLM methods and gives best values for the PSNR measure which means that there is no difference between the original and the stego-images which implies a good visual quality of the generated stego-image. Also the experimental results show that this method hides more information.

B. Transform Domain Embedding

LSB embedding techniques are the easiest way to insert the secret message into the cover because the message bits are inserted directly to the image pixels bits but as discussed they are highly vulnerable to be detected by steganalysis. By the rapid development of information technology and the increased need for more secure steganography algorithms Transform domain embedding algorithms has been proposed [26-32] trying to provide more robustness against attacks than their ancestors (spatial domain embedding methods). Transform domain methods hides the messages in significant areas of the cover image to produce more efficient stego images. It manipulates the image indirectly by various transformation techniques; the most popular of these techniques are discussed below:

Discrete Cosine Transformation (DCT): it is the most popular in the transform domain because that the DCT based image format (JPEG) are widely available and it is the common output of digital cameras. DCT transforms successive 8*8 pixel blocks of the image from spatial domain to 64DCT coefficients, and after calculating the coefficients they are altered to embed the secret message into them (i.e. modifying the LSB's of each coefficient) [26]. Reference [27] proposes a steganography method based on Integer DCT and affine transformation and the experimental results shows that this

method provides a stego images that is visually and statistically undetectable.

Discrete Wavelet Transformation (DWT): the standard techniques of embedding using LSB still applied here but the difference is that the secret message bits are embedded into the wavelet coefficient bits (LSB) rather than changing bits of the actual pixel bits.

Reference [31] provide a method that retains the integrity of the wavelet coefficients even at high capacity embedding which was achieved by estimating the capacity of each DWT block and applying the embedding process to the whole block instead of the bit-planes. This method also guarantees that no noisy bit-plane left unused. Therefore achieving more capacity than other methods as proved in the experimental results.

Reference [32] provides an algorithm to hide text in any colored image of any size using wavelet transform. The experimental results prove that this algorithm improves the image quality and imperceptibility.

Integer Wavelet Transformation (IWT): embeds the secret image in frequency domain of cover image with high matching quality [29].

Reference [29] used the IWT to transform both cover and secret images from spatial domain to frequency domain and then an assignment algorithm is used to select the best matching between blocks for embedding. There experimental results show that this method provides high robustness against different attacks and gives better PSNR.

Reference [30] takes the advantage of iterative blending to propose a novel self-adaptive image steganography scheme based on IWT. This scheme provides excellent properties of invisibility and robustness also the capacity is increased by the use of iterative blending techniques as shown on their experimental results.

C. Hybrid Approaches

In this type of steganography spatial and transform domains may be combined [33] with each other, also one or both of them maybe combined with other optimization algorithms or heuristic approaches like Genetic Algorithms (GA) [34] and [35], or may be combined with other secure communication techniques like cryptography [36] and [37]. Aiming at producing a new schemes that will provide better results than if they used alone.

Reference [33] combines both spatial and transforms domains to give greater security. They use LSB and DWT. The experimental results show that the proposed scheme provides high embedding capacity as well as high level of security.

Reference [35] proposed a method that combines spatial domain steganography and genetic algorithm. The results shows that this method provide high embedding capacity and enhances the PSNR measure of the stego image, also [34] also combines genetic algorithm with the wavelet transformation scheme which also increase the capacity and imperceptibility of the stego image.

Both [36] and [37] used visual cryptography along with image steganography to achieve more security.

IV. PERFORMANCE SPECIFICATION OF IMAGE STEGANOGRAPHY FACTORS

The ultimate goal for all researchers in the field of image steganography is to find an algorithm that provide high embedding capacity that is also highly secure and imperceptible; these concepts are discussed earlier in the introduction.

For the stenographer it is important to show and analyze the relation between the three factors to make them working together. This relation can be presented by the steganography triangle (Fig. 3).

Fig. 3, represents a balance triangle each of its ribs specifies a factor associated with a steganographic method. So that for instance in order to improve capacity, you sacrifice security. It makes scenes that the more embedding in an image the more probability that an observer will notice the degradation and suspect something is out of place. It is obvious that improving one factor will affect the other factors so that any steganography method must take care of the three factors at all times trying to keep the triangle as balanced as possible you have to change the other two elements.

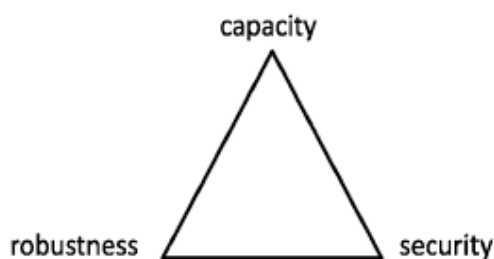


Fig. 3. The Steganography Triangle.[40]

V. CRITICAL ANALYSIS

A critical comparison and analysis of the main researches that have been discussed from the three image steganography domains is presented in Table I.

TABLE I. CRITICAL ANALYSIS OF STEGANOGRAPHY LETRETURE

Lit. Ref	Analysis			
	Domain	Technique	Advantage	Disadvantage
[1]	Spatial	ELSB	Less distortion	Robustness
[2]	Spatial	LSB & Edge Detection	High payload, Image quality	Robustness
[3]	Spatial	LSB & Edge Detection	High Security	Robustness

Lit. Ref	Analysis			
	Domain	Technique	Advantage	Disadvantage
[4]	Spatial	LSB & Edge Detection	High payload	Robustness, Imperceptibility
[5]	Spatial	Edge Detection	High payload Imperceptibility Good Robustness PSNR	Compression
[11]	Spatial	LSBMR	Decrease required modifications to the image	Robustness, Poor visual quality, Low security
[12]	Spatial	LSBMR & Edge adaptive	Preserve image statistical and visual quality	Detected by statistical steganalysis
[13]	Spatial	Block-Based edge adaptive	High security	Robustness, Payload
[15]	Spatial	EA-LSBMR	High security, Imperceptibility	Robustness
[23]	Spatial	MBP	High security, against classical, steganalysis	Robustness, Compression
[24]	Spatial	MBNS	Imperceptibility	Robustness
[25]	Spatial	LSB & GLM	High security, Imperceptibility, High payload	Robustness
[27]	Transform	DCT	Imperceptibility, Robustness	Payload
[28]	Transform	DFT	Imperceptibility Robustness	Payload
[29]	Transform	IWT	High Robustness, High PSNR	Payload
[30]	Transform	IWT & Iterative blending	Imperceptibility Robustness Good embedding capacity	Payload still less than LSB
[31]	Transform	DWT	Good embedding capacity	Payload still less than LSB
[32]	Transform	DWT	Imperceptibility Robustness	Payload
[33]	Spatial & Frequency	LSB & DWT	High security, High embedding capacity	Robustness
[34], [35]	Hybrid	DWT & Genetic	High embedding capacity, Imperceptibility	Robustness
[36], [37]	Hybrid	LSB & Visual	High security	Robustness

Lit. Ref	Analysis			
	Domain	Technique	Advantage	Disadvantage
		cryptogra- phy		

VI. CONCLUSION

This article discussed the digital image steganography and steganalysis. Steganography is the science that involves communicating secret data in an appropriate multimedia carrier, e.g., image, audio, and video files. While Steganalysis is the science of attacking steganography, also is not the focus of this survey but nonetheless it was briefly discussed. Being different from cryptography technique, steganography provides an approach for solving the increased security problems, and it has become the new research hotspot in the field of international information security. Robustness, imperceptibility, and hiding capacity are the three main evaluation standard of steganography technique [17]. The inevitable conflict between the three steganography factors was discussed. Take imperceptibility and hiding capacity as the two factors for example, the larger hiding capacity, which means more modifications will be implemented with cover image, which implies lower imperceptibility.

The main image steganography domains is Spatial domain and Transform domain also the Hybrid approaches have been discussed concluding that the spatial domain schemes provides high payload embedding and good visual quality but it is simple and highly vulnerable to security attacks specially the statistical steganalysis. On the other hand the transform domain schemes have many advantages, including its persistence against statistical attacks along with strong robustness however; they usually hide less capacity of secret information.

The hybrid approaches also provide some security and capacity enhancements but still in its beginning and need more and more research.

Finally a question pops to our minds, when guaranteeing excellent imperceptibility, how to increase the hiding capacity? or vice versa. More in-depth research is needed.

REFERENCES

[1] Shilpa Gupta, Geeta Gujral and Neha Aggarwal “Enhanced least significant bit algorithm for image steganography” IJCEM International Journal of Computational Engineering & Management, Vol. 15 Issue 4, July 2012.

[2] Chen, W.-J., C.-C. Chang, and T. Le “High payload steganography mechanism using hybrid edge detector” Expert Systems with applications, Vol. 37, pp. 3292-3301, 2010.

[3] Jain, N., S. Meshram, and S. Dube, “Image steganography using LSB and edge-detection technique” International Journal of Soft Computing and Engineering (IJSCE), 2012.

[4] Ioannidou, A., S.T. Halkidis, and G. Stephanides “A novel technique for image steganography based on a high payload method and edge detection” Expert Systems with Applications, Vol. 39, pp. 11517-11524, 2012.

[5] Shahzad Alam, Vipin Kumar, Waseem A Siddiqui and Musheer Ahmad “Key dependent image steganography using edge detection” Fourth

International Conference on Advanced Computing & Communication Technologies (ACCT), 2014.

[6] Fridrich, J., M. Goljan, and R. Du. “Reliable detection of LSB steganography in color and grayscale images.” Proceedings of the workshop on Multimedia and security: new challenges. 2001.

[7] Lu, P., et al. “An improved sample pairs method for detection of LSB embedding in Information Hiding”. 2005.

[8] Luo, X., B. Liu, and F. Liu “ Improved RS method for detection of LSB steganography” Computational Science and Its Applications-ICCSA , pp. 508-516, 2005.

[9] Ker, A.D. “A general framework for structural steganalysis of LSB replacement in Information Hiding”. 2005.

[10] Dumitrescu, S. and X. Wu “A new framework of LSB steganalysis of digital media” Signal Processing, IEEE Transactions on, Vol. 53, pp. 3936-3947, 2005.

[11] Mielikainen, J., “LSB matching revisited” IEEE Signal Processing Letters, Vol. 13, pp. 285-287, 2006.

[12] Luo, W., F. Huang, and J. Huang “Edge adaptive image steganography based on LSB matching revisited” Information Forensics and Security, Vol. 5, pp. 201-214, 2010.

[13] Huang, W., Y. Zhao, and R.-R. Ni “Block based adaptive image steganography using LSB matching revisited” J. Elec. Sci. Tech, Vol. 9 , pp. 291-296, 2011.

[14] Tan, S. and B. Li, “Targeted steganalysis of edge adaptive image steganography based on LSB matching revisited using B-spline fitting” IEEE Signal Processing Letters, Vol. 19, pp. 336-339, 2012.

[15] Huang, F., Y. Zhong, and J. Huang, “Improved algorithm of edge adaptive image steganography based on LSB matching revisited algorithm” n Digital-Forensics and Watermarking, pp. 19-31, 2014.

[16] Dr. Ekta Walia , Payal Jain , Navdeep “An Analysis of LSB & DCT based Steganography” Vol. 10, April 2010.

[17] Niels Provos and Peter Honeyman “Hide and seek: An introduction to steganography” IEEE Security and Privacy, vol. 1, no.3, pp. 32-44, 2003.

[18] T Mrkel, JHP Eloff and MS Olivier “An Overview of Image Steganography” proceedings of the fifth annual Information Security South Africa Conference , 2005.

[19] Hardik Patel, Preeti Dave “Steganography technique based on DCT coefficients” International Journal of Engineering Research and Applications, Vol. 2, pp.713-717, Jan-Feb 2012.

[20] Robert Krenn, “Steganography and steganalysis” Internet Publication, March 2004. Available at: <http://www.krenn.nl/univ/cry/steg/article.pdf>. Accessed on December 14, 2014.

[21] Souvik Bhattacharyya, Indradip Banerjee and Gautam Sanyal “A survey of steganography and steganalysis technique in image, text, audio and video as cover carrier” Journal of Global Research in Computer Science Vol. 2, April 2011.

[22] Bin Li, Junhui He, Jiwu Huang and Yun Qing Shi, “A survey on image steganography and steganalysis,” Journal of Information Hiding and Multimedia Signal Processing, vol. 2, April 2011.

[23] B.C. Nguyen, S.M. Yoon et H.-K. Lee “Multi bit plane image steganography” Digital Watermarking, 5th International Workshop, IWDW, Vol. 4283, Novembre 2006.

[24] Xinpeng Zhang and Shuozhong Wang “Steganography using multiple-base notational system and human vision sensitivity” IEEE Signal Processing Letters, Vol. 12, 2005.

[25] Ahmad T. Al-Taani. and Abdullah M. AL-Issa.” A novel steganographic method for gray-level images” International Journal of Computer, Information, and Systems Science, and Engineering, Vol. 3, 2009.

[26] Anu, rekha, Praveen “Digital image steganography” International Journal of Computer Science & Informatics, Vol. 1, 2011.

[27] Xianhua Song, Shen Wang and Xiamu Niu “An Integer DCT and Affine Transformation Based Image Steganography Method” IEEE Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, 2012.

- [28] Ashish Soni, Jitendra Jain and Rakesh Roshan “Image steganography using discrete fractional fourier transform” IEEE International Conference on Intelligent Systems and Signal Processing (ISSP) 2013.
- [29] Neda Raftari , Amir Masoud and Eftekhari Moghadam “Digital image steganography based on integer wavelet transform and assignment algorithm” IEEE Sixth Asia Modelling Symposium, 2012.
- [30] Peipei Liu, Chuan Chen, Liangquan Ge, and Yaoyao Luo “Efficient self-adaptive image steganography scheme based on iterative blending and integer wavelet transform” Springer Lecture Notes in Electrical Engineering, 2014.
- [31] Saeed Sarreshtedari and Shahrokh Ghaemmaghami “High capacity image steganography in wavelet domain” IEEE CCNC, 2010 proceedings.
- [32] Saddaf Rubab and M. Younus “Improved image steganography technique for colored images using wavelet transform” International Journal of Computer Applications, Vol. 39 ,pp. 0975 – 8887, February 2012.
- [33] Saurabh V. Joshi , Ajinkya A. Bokil, Nikhil A. Jain and Deepali Koshti “Image steganography combination of spatial and frequency domain” International journal of computer applications. Vol. 53, September 2012.
- [34] Elham Ghasemi, Jamshid Shanbehzadeh and Nima Fassihi “High capacity image steganography using wavelet transform and genetic algorithm” International MultiConference of Engineers and Computer Scinetests. Vol. 1, 2011.
- [35] Hamidreza Rashidy Kanan and Bahram Nazeri “A novel image steganography scheme with high embedding capacity and tunable visual image quality based on a genetic algorithm” Expert Systems with Applications, Vol. 41, pp. 6123–6130, 2014.
- [36] D. R. L. Prasanna, L. Jani Anbarasi and M. Jenila Vincent “A Novel Approach for Secret Data Transfer using Image Steganography and Visual Cryptography” ICCCS’11, February 2011.
- [37] Piyush Marwaha and Paresh Marwaha “Visual Cryptographic Steganography In Images” Second International conference on Computing, Communication and Networking Technologies. 2010.
- [38] Cheddad, A., Condell, J., Curran, K., & Mc Kevitt, P. “Digital image steganography: Survey and analysis of current methods” Signal processing, Vol. 90(3), pp. 727-752, 2010.
- [39] Kaur, S., Kaur, A., & Singh, K. “A survey of image steganography” IJRECE, Vol. 2(3), pp. 102-105, 2014.
- [40] Huayong, G., Mingsheng, H., & Qian, W. “Steganography and Steganalysis based on digital image” IEEE Image and Signal Processing (CISP) 4th International Congress, Vol. 1, pp. 252-255). October 2011.

COMPUTER VISION APPLIED TO ROAD LINES RECOGNITION USING MACHINE LEARNING

C. H. Rodríguez-Garavito, A. Ponz, F. García, A. de la Escalera and J.M. Armingol.

Intelligent Systems Lab.
Universidad Carlos III de Madrid, UC3M.
Madrid, Spain.

cesarhernan.rodriguez@alumnos.uc3m.es , {apv, fegarcia, escalera, armingol}@ing.ucm3.es

C. H. Rodríguez-Garavito.
Automation engineering department.
Universidad de La Salle, Bogotá Colombia.
cerodriguez@unisalle.edu.co

Abstract— According to the Department for Transport statistics in UK, around 100.000 accidents were reported in 2013 [13], and almost 25% of them were related to impairment or distraction factors. Advanced Driver Assistance Systems (ADAS) are a powerful tool for road safety that can help to mitigate this problem. This paper presents a robust road lane detection and classification algorithm, one of the most important tasks in ADAS. This paper describes a road line detection algorithm based on a segmentation algorithm designed according to the constraints defined in the legal regulation for road marks. Later, pairs of lines, separated a fixed distance, are searched in the bird view of the road image. The bird view transformation is applied to the captured images, using the extrinsic parameters estimation algorithm reported in [10]. After the extraction of the road lines profiles, they are characterized using a specifically designed descriptor based on both space and frequency values. The descriptors are used in the supervised training of a Support Vector Machines classifier, whose performance is compared against the previous version of the module, a heuristic based approach. The performed tests showed a considerable increase of the system performance using the SVM approach, in comparison with the previous heuristic approach.

Keywords— Road Line Segmentation; Road Line Classification; Support Vector Machine, Bird Eye View

I. INTRODUCTION

Nowadays, interest in Advanced Driving Assistance Systems (ADAS) is growing due to their contribution to effectively reduce human error related accidents; therefore, they have become a differentiating element in the automotive market.

One of the key points in ADAS applications are automatic and intelligent algorithms and strategies such as machine learning. Among the multiple tasks that might take advantage of these useful techniques are: Detection and identification of road signs [1], vehicle [2] and pedestrian [3] detection, estimation of driver distraction [4], and environment interpretation and understanding [5].

The present work is focus on the extraction and detection of road lane delimitation lines. This task must be reliable in order to build a complete model of the road environment, that is, determine the type of road, number of lanes, type of lanes, and vehicle position with reference to the current lane, among many other characteristics.

Several works have already addressed this issue on the basis of different techniques with promising results: in [6] road lines are classified as solid or dashed, by analyzing the gaps

between measurement points, which are predicted throughout a projected model lane and a complex kalman filter. In [7], line boundaries detection is performed based on a priori information about its position in the processed image; a ROI at the initial position of each boundary is used to get a statistic intensity model and built a binary path for each line. Next, a cascade classifier is applied, and the sub-classifiers are tuned for dashed, dashed-solid, solid-dashed, single-solid and double-solid line patterns, they are based on measure like fraction of marking-related pixels and temporal autocorrelation. Another approach using off-line information of the road curvature is proposed in [8], where a priori information, together with external sensors measurements from GPS and gyroscope, allow tracking model estimation by the use of an AMF (Approximated Median Filter).

This paper presents the Road Lane Classification module of the Ivvi 2.0 project (Intelligent Vehicle based on Visual Information) [12]. Its goal is to automatically detect the position, type, and number of the road lanes with a stereo on-board camera inside of the Intelligent System Lab at Carlos III University experimental vehicle. In this work, three types of lane boundaries are considered, namely: solid, dashed and

merge, using classification by heuristic classifier [9] and support vector machine, SVM classifier.

The article is divided into the following sections: Section II explains how lane boundary lines are segmented. Section III focuses on the algorithms for line description and classification. Section IV describes the performance tests executed for the classifiers, and Section V presents some conclusions for the work.

II. SEGMENTATION OF IN-ROAD LANE SEPARATOR LINES

The process starts with the capture of a pair of stereo images, from where a three-dimensional representation of the road environment is built. This representation is called Point Cloud (PC), and is a collection of points in the space referred to a system of coordinates located in the reference camera of the stereo pair. As described in [10], it is possible to obtain the parametric definition of the most populated plane in the PC. Some information obtained from this plane determines the spatial position of the camera that records the images. The parameters that define this spatial position are called “extrinsic camera parameters”, and together with the intrinsic camera parameters, that is, sensor and optical system specifications, allows the conversion of the actual perspective of the captured scene into a more convenient perspective for road lines detection. In this new perspective, called “bird view perspective” or “upper view”, the parallel geometry between road lines in the real world is visually preserved. This perspective transformation or homograph is shown in Equation

$$s * {}^cP = {}^cP_s = {}^cH_{bv} {}^{bv}P \tag{1}$$

where a relation is established between every pixel position ${}^{bv}P_i$ and pixels cP_i in the original image, s is an scaling factor which appears as the third component of the point resulting from the homographic transformation, cP_s .

The matrix ${}^cH_{bv}$ is decomposed as seen in Equation (2), details are shown in Table 1.

$${}^cH_{bv} = K * {}^cT_m * {}^mT_{bv} * K^{-1} \tag{2}$$

TABLE I. PARAMETRIC DESCRIPTION OF A HOMOGRAPHIC TRANSFORMATION.

Parametric Matrix	Meaning
K	Intrinsics matrix
cT_m	Spatial transformation from road plane into camera position
${}^mT_{bv}$	Spatial transformation from bird view position into road plane

The result of the transformation can be seen in Fig. 1.



Fig. 1. Homographic transformation from road image into bird view and road lines detection.

After the homographic transformation, a detection of the pixels meeting the characteristics of a road line is performed. For the present approach, these characteristics were defined on the basis of variations of the intensity gradient, direction gradient, and the road line width, in order to create a road line mask as shown in Fig. 2.



Fig. 2. Road line mask in bird view perspective.

A. Road lines and lanes detection

Using the Hough Transform, line detection is performed in the bird view image over road line mask. The direction of the lines belonging to the road is not known a priori, so the assumption that the angle θ_R of the searched lanes will match the angle with the highest magnitude in a histogram of orientations computed over the total of detected lines will be made.

From the set of detected lines, the pairs of lines separated a specific distance (i.e. lane width \pm road mark width threshold) parallel to each other and whose orientation matches the road angle, θ_R , are identified. These characteristic were based on Spain’s IC 8.2 standard [11]. The filtered lanes are not necessarily consecutive, but may be overlapping as shown in Fig. 3.

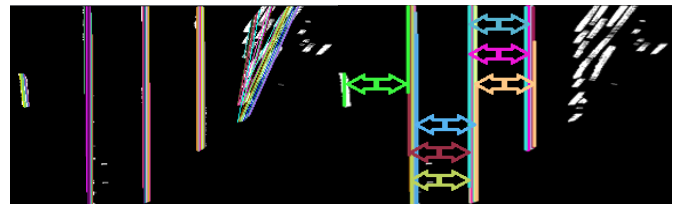


Fig. 3. Lines detected with Hough Transform and non-consecutive lanes segmented.

B. Identification of lanes on the road in adjacent fashion

In order to create a spatially ordered structure, i.e. a road with consecutive lanes, the detected lanes are chained using a

separating line identifier, thus it is possible to find their order in the road.

Two successive ordering processes are performed. A first ordering is made, starting at a random lane. Initially, one of the two directions, $\theta_R \pm \pi$ in which adjacent lanes are labeled from an initial lane. Once the search in one direction is finished, a new search is performed, starting from the first lane in the opposite direction.

The lanes obtained from the search are stored in consecutive order. Each one is composed of a pair of lines matching the aforementioned separation constraints. Information from common boundaries is fused and unique lines between lanes are created as shown in Fig. 4.



Fig. 4. Lane Boundaries defined by fusing spatially adjacent lanes.

III. LINES EXTRACTION AND CLASSIFICATION

Road lines classification starts with the extraction of detected line profiles, $f(\text{pix})$. Profiles are vectors of binary values whose size matches the length of the detected line. The array is initialized by scanning the positions of the detected line on the road lines mask. An example of a detected line profile can be seen in Fig. 5.

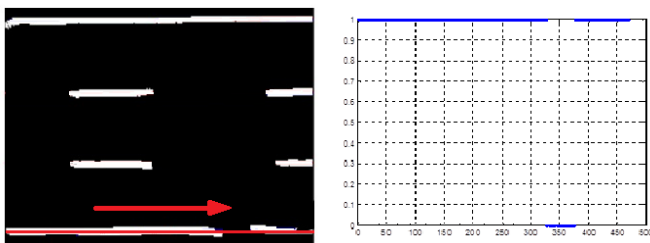


Fig. 5. Detected line profile, $f(\text{pix})$.

The detected lines are obtained from the road lines mask, which may contain noise and spurious elements. Thus, a fuzzy logic strategy called center of gravity, is employed in order to defuzzify an end-of-line profile into a corrected concrete value.

C. Adjustment by center of gravity

The position of each end-of-line is corrected by computing the center of gravity of its profile, which is taken in the direction perpendicular to the line checked, in a neighborhood of two road line widths, as depicted in Fig. 6.

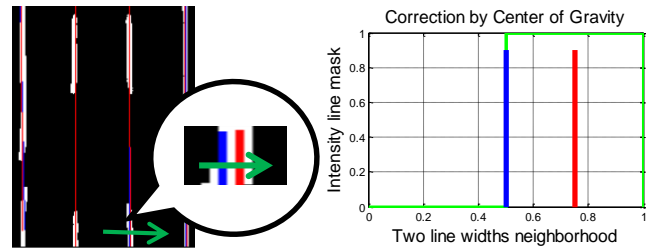


Fig. 6. Correction by center of gravity of the end of line coordinates for each lane in the road ordered structure.

D. Descriptor

The next step in the machine learning based strategy for road lines classification is the unequivocal description of the elements to classify, that is, to find a representation that unambiguously identifies all the possible elements to classify. Two different elements can't share the same descriptor; furthermore, the descriptors must be as small as possible in order to allow efficient training and prediction processes.

The descriptor chosen in this case is based on a mix of features in both spatial and frequency domain. The two first features are the mean value of the line profile, and its length, in meters. The remaining features correspond to the frequencies of the n first power peaks, in descending order. The selected features take into account the difference in the spatial frequency representation of the different line types, as seen in Fig. 7.


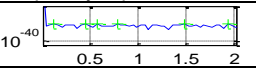

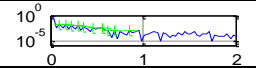

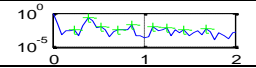
Line type	Spatial representation	Frequency representation
Solid	 $=0.1$	
Dashed	 $=0.1$	
Merge	 $=0.5$	

Fig. 7. Line profiles characterization in spatial and frequency domains.

Fig. 8 shows the power spectrum $F(f(\text{pix}))$ for a detected line, $f(\text{pix})$. Two different methods were developed for line classification: A heuristic method, based on the behavior of the intensity peaks in the descriptor, and a second method based on Support Vector Machines, a classification machine learning algorithm.

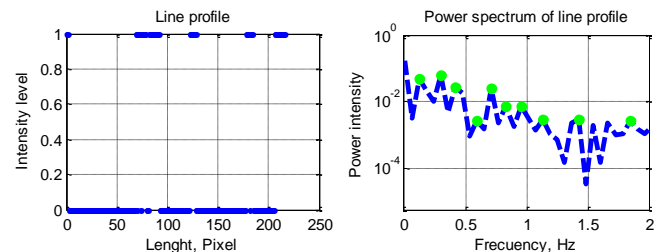


Fig. 8. Power spectrum F , for a detected line, $f(\text{pix})$.

E. Heuristic classifier

In the case of the heuristic classifier, the decision rules are obtained by analyzing a great collection of sampled lines, along with their corresponding space and frequency spectrum, as shown in Fig. 9 and Fig. 10.

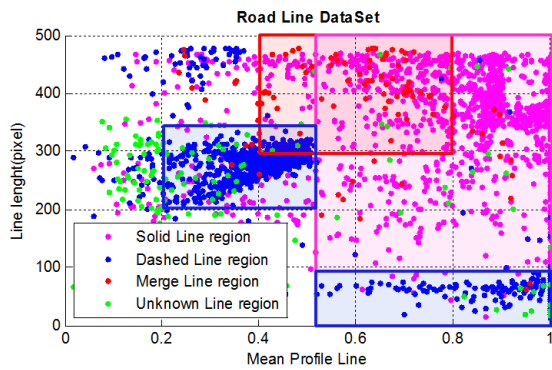


Fig. 9. Feature Space for descriptor lines.

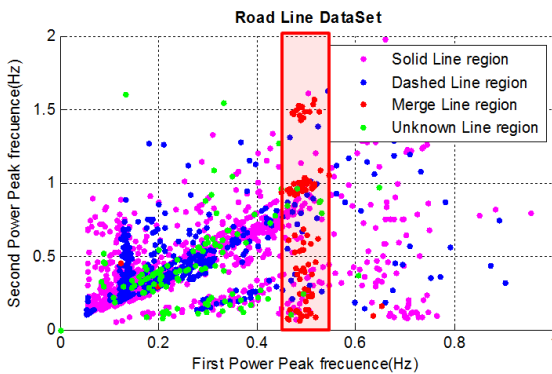


Fig. 10. Feature frequency for descriptor lines.

The decision algorithm is a sequence of three binary classifiers, C_1 , C_2 y C_3 . shown in Fig. 11.

- ω_1 : The distance from 0.5 Hz to some of the 3 first peaks of F is smaller than 0.05 Hz?
- ω_2 : The mean of f is greater than 0.4 and smaller than 0.8?
- ω_3 : The length of f is greater than the three lengths of a dashed line segment?
- ω_4 : The mean of f is greater than 0.52?
- ω_5 : The length of f is greater than the length of a dashed line segment?
- ω_6 : The mean of f is greater than 0.2 and smaller than 0.52?
- ω_7 : The length of f is greater than the 2 lengths of a dashed line segment and smaller than 3.3 lengths of a dashed line segment?

- ω_8 : The length of f is smaller than the length of a dashed line segment?

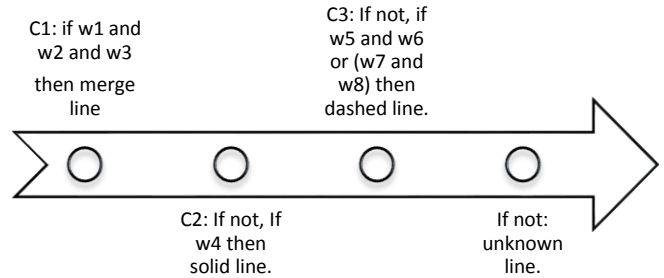


Fig. 11. Decision algorithm for heuristic classification.

F. SVM classifier

SVM, the machine learning based method, uses the same descriptor, and associates each line type to a different class. The learning and prediction approaches are shown in Fig. 12.

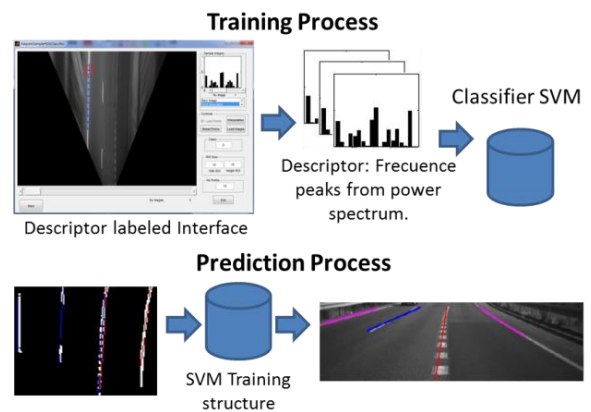


Fig. 12. Stages of the machine learning classification process.

Two different processes were necessary to develop the SVM classification algorithm. The first one refers to the database creation, populated with an adequate amount of manually labeled descriptors, some of them are later used as a ground truth. This information database also provides the supervised learning examples that iteratively adjust an optimal parametric SVM structure. The classifier is then implemented using a radial base kernel and a multiclass strategy. Four binary classifiers are trained, one per class.

After the training process, the trained SVM structures are stored for recovery in the prediction stage of new non-trained descriptors. The detection is made by checking how close is an input descriptor to each classification region given by a different classifier. The closest region to the descriptor determines the selected class. A visual representation of the classification regions, using line profile mean and first power peak frequency as classification features, is shown in Fig. 13.

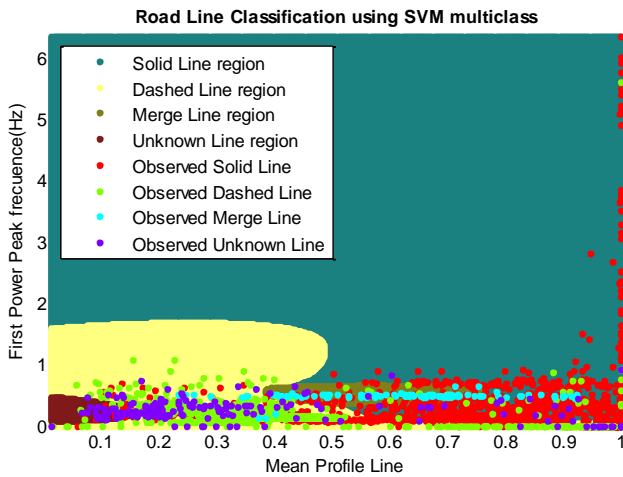


Fig. 13. Road lines classification using multiclass SVM.

IV. TESTS AND RESULTS

Tests of the road lines classifiers are made over a database of 1000 images. Based on them, a collection of 4100 manually labeled descriptor-class pairs were made. 80% of data was used for training and 20% of data was used for validation.

Table 2 depicts the confusion matrix associated to each classification process with SVM descriptor. Table 3 shows the confusion matrix associated to heuristic classification. These are the results of prediction versus ground truth comparisons over the validation data.

TABLE II. CONFUSION MATRIX SVM CLASSIFIER.

	Solid	Dashed	Merge	Unkown
Solid	298	24	6	13
Dashed	17	375	0	3
Merge	5	2	25	0
Unkown	7	10	1	19

TABLE III. CONFUSION MATRIX HEURISTIC CLASSIFIER.

	Solid	Dashed	Merge	Unkown
Solid	193	78	40	30
Dashed	3	334	13	45
Merge	2	2	16	12
Unkown	2	6	5	24

Finally, based on the classification results, classification methods performances can be measured using percentage of true positives per class, as shown in Fig. 12. and Table 4.

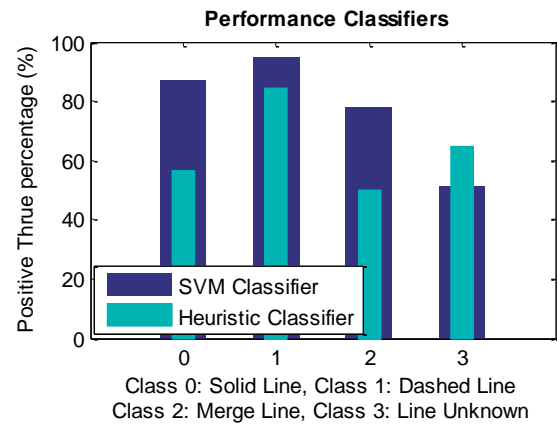


Fig. 14. Performance of the analyzed classifiers: Heuristic and SVM.

TABLE IV. PERCENTAGE RECOGNITION LINE CLASS BY CALSSIFICATION METHOD.

Method	Percentage Recognition Line Class			
	Solid	Dashed	Merge	Unknown
Heuristic	0.566	0.8456	0.5	0.6486
SVM	0.8739	0.9494	0.7813	0.5135

V. CONCLUSIONS

An improved version of the Ivvl 2.0 road lines detection and classification module is presented. The road lanes segmentation procedure shown has proven to be a robust method, invariant both to high shapes changes and to occlusions. This segmentation stage is appropriate for further lane tracking.

The main stage of the work, that is, road line classification using SVM over a 1000 images collection, shows a significant improvement in true positive detections. Success rates increased a 31% for solid lines, 10% for dashed lines, and 28% for merge lines. This classification strategy is a systematic method that allows for the inclusion of further images characteristics, such as changes in illumination conditions or new road signs.

The remarkable detection line results obtained in the aforementioned tests were achieved based on single frame detection. Further steps will be focused on developing a tracking method that will help to overcome difficult situations and exceptional misclassifications by adding time consistency.

ACKNOWLEDGMENTS

This work was supported by automation engineering department from de La Salle University, Bogotá-Colombia; Administrative Department of Science, Technology and Innovation (COLCIENCIAS), Bogotá-Colombia and the Spanish Government through the CICYT projects (TRA2013-48314-C3-1-R) and (TRA2011-29454-C03-02) and Comunidad de Madrid through SEGVAUTO-TRIES (S2013/MIT-2713).

REFERENCES

- [1] Maldonado-Bascon, S., Lafuente-Arroyo, S., Gil-Jimenez, P., Gomez-Moreno, H., & Lopez-Ferreras, F. (2007). Road-sign detection and recognition based on support vector machines. *Intelligent Transportation Systems, IEEE Transactions On*, 8(2), 264-278. doi:10.1109/TITS.2007.895311
- [2] Zezhi Chen, Pears, N., Freeman, M., & Austin, J. (2009). Road vehicle classification using support vector machines. *Intelligent Computing and Intelligent Systems, 2009. ICIS 2009. IEEE International Conference On*, 4, 214-218. doi:10.1109/ICICISYS.2009.5357707
- [3] Olmeda, D., de la Escalera, A., & Armingol, J. M. (2009). Detection and tracking of pedestrians in infrared images. *Signals, Circuits and Systems (SCS), 2009 3rd International Conference On*, 1-6. doi:10.1109/ICSCS.2009.5412297
- [4] Yulan Liang, Reyes, M. L., & Lee, J. D. (2007). Real-time detection of driver cognitive distraction using support vector machines. *Intelligent Transportation Systems, IEEE Transactions On*, 8(2), 340-350. doi:10.1109/TITS.2007.895298
- [5] Zhang, H., Hou, D., & Zhou, Z. (2005). A novel lane detection algorithm based on support vector machine. *Progress in Electromagnetics Research Symposium, Hangzhou, China*,
- [6] Risack, R., Klausmann, P., Krüger, W., & Enkelmann, W. (1998). Robust lane recognition embedded in a real-time driver assistance system. In *Proc. IEEE IV*,
- [7] Paula, M. B. d., & Jung, C. R. (2013). Real-time detection and classification of road lane markings. *Graphics, Patterns and Images (SIBGRAPI), 2013 26th SIBGRAPI-Conference On*, 83-90.
- [8] Maeda, T., Hu, Z., Wang, C., & Uchimura, K. (2008). High-speed lane detection for road geometry estimation and vehicle localization. *SICE Annual Conference, 2008*, 860-865.
- [9] Collado, J. M., Hilario, C., De La Escalera, A., & Armingol, J. M. (2006). Adaptive road lanes detection and classification. *Advanced Concepts for Intelligent Vision Systems*, 1151-1162.
- [10] Rodriguez-Garavito, C., Ponz, A., Garcia, F., Martín, D., de la Escalera, A., & Armingol, J. (2014). Automatic laser and camera extrinsic calibration for data fusion using road plane. *Information Fusion (FUSION), 2014 17th International Conference On*, 1-6.
- [11] Espanya., D. G. d. C. (1991). *Marcas viales : Norma de carreteras 8.2-IC*. Madrid: MOPU. Secretaría General Técnica. Centro de Publicaciones.
- [12] Martín, D., García, F., Musleh, B., Olmeda, D., Peláez, G., Marín, P., Armingol, J. M. (2014). IVVI 2.0: An intelligent vehicle based on computational perception. *Expert Systems with Applications*, 41(17), 7927-7944. doi:<http://dx.doi.org/10.1016/j.eswa.2014.07.002>
- [13] DfT. (2013). *Reported road casualties Great Britain: 2013. Annual Report*. London: The Stationery Office.

Combining ICA Representations for Recognizing Faces

Ashraf Y. A. Maghari

Faculty of Information Technology
Islamic University of Gaza
Gaza, Palestine
amaghari@iugaza.edu.ps

Abstract— Independent Component Analysis (ICA) is a generalization of Principal Component Analysis (PCA), and it looks for components that are both statistically independent and non-Gaussian. ICA is sensitive to high-order statistic and it expected to outperform PCA in finding better basis images. Moreover, with face recognition, high-order relationships among pixels may have more important information than those of pairwise relationships on which base images found by PCA depend. Two different representations can be applied by ICA; ICA architecture I and ICA architecture II. A new classifier that combines the two ICA architectures is proposed for face recognition. By the new classifier, the similarity measure vector was employed in which the similarity measure vectors for both ICA representations were resorted in descending order and then integrated by merging the corresponding values of each vector. The new classifier was performed on face images in the AR Face Database. Cumulative Match Characteristic was taken as a measure for evaluating the performance of the new classifier with illumination variation, expression, and Occlusion. The proposed classifier outperforms both ICA architectures in all cases especially in later ranks.

Keywords— ICA; PCA; face recognition; ICA representations

I. INTRODUCTION

Human face recognition is a very important issue in the field of Computer Vision. It has gained a lot of attention during the last decades [1]. Current 2D face recognition systems can achieve good performance in constrained environments. However, they still encounter difficulties in handling large amounts of facial variations [2]. Extensive research is ongoing to make face recognition systems robust to typical operational environments where uncertainties such as occlusion, illumination and other variations are common [3]. In the last decades, a number of biometric face recognition algorithms have been proposed by computer scientist, neuroscientists and psychologist's efforts. The computer scientists seek to develop methods for face recognition whereas the psychologists and neuroscientists work on biological perception of human face recognition process i.e. face recognition is done holistically or feature analysis etc. [4]. There are two important matters in face recognition algorithms, feature representation and classification based on features. Based on feature representation; face recognition methods can be classified into two groups i.e. face and constituent. Face-based methods (appearance based technique) use raw information face pixel, whereas constituent based approaches use the relationships between face features i.e. nose, lips, and eyes. Compared to face-based methods, the constituent-based methods are more flexible but the performance depend on features [4].

Bio-inspired evolutionary search was presented in [4] where a constituent-based method was employed with fixed fiducial points extracted from the face image. GA was then used to search best feature combination that gives minimum training error. However, in this work, we used the appearance based technique so as the image is considered as 2D pattern. Among appearance based representation, PCA based method is one of the most powerful methods successfully applied in face recognition [4]. Principle Component Analysis (PCA) is a popular unsupervised statistical method used for dimension reduction and face recognition. The goal of PCA is to find a set of basis images so that the PCA coefficients are linearly independent. The performance of PCA depends on the task statement, the subspace distance metric, and the number of subspace dimensions retained. Independent Component Analysis (ICA) can be seen as a generational of PCA and its basic idea is to represent a set of random variables using basis functions, where the components are statistically independent or as independent as possible [5].

There are many algorithms used for performing ICA [6], [7]. In this study, the information maximization learning algorithm developed by [8], [9] is employed. The algorithm was developed from the principle of optimal information transfer in neural networks with sigmoidal transfer functions. It has been shown that this algorithm has proven successful for separating electroencephalogram (EEG) signals [10] and functional magnetic resonance imaging (fMRI) signals [11]. The non-locality of the learning algorithm is interesting when

biological significance of the learned filters is considered [9]. However, the biological plausibility of the algorithm is limited when the learning rule is nonlocal. The performance of ICA depends on the task, the algorithm used to approximate ICA, and the number of subspace dimensions retained. There are two different approaches of ICA to face recognition. ICA can be applied so as to treat images as random variables and pixels as observations, or to treat pixels as random variables and images as observations. In consistence with [12], [13], we refer to these two alternatives as ICA architecture I and architecture II, respectively.

The two architectures of ICA were performed by Bartlett et al. [13] on face images in the FERET database. They developed a classifier that combine the ICA representations to give better performance in comparison with the two architectures. The combined classifier was employed in which the similarity between the test image and the gallery image has been defined as the summation of the similarity measure of both ICA1 and ICA2 (ICA1+2).

In this study, we developed a new classifier that combine both ICA1 and ICA2 in different way. The classifier was adopted in which the similarity measure vectors for both ICA1 and ICA2 were resorted in descending order and then integrated by merging the corresponding values of the two vectors to reconstruct a new similarity measure vector. AR database has been used to train and test the new combined classifier.

This paper is organized as follows: Section II gives brief introduction on ICA. In Section III, the proposed classifier will be demonstrated, whereas in Section IV the Experiment results and discussion are reported. In the last section, the paper is summarized.

II. INDEPENDENT COMPONENT ANALYSIS (ICA)

ICA is a generalized form of Principal Component Analysis (PCA), which is able to separate independent sources linearly mixed in several signals. The Information Maximization algorithm proposed by Bell and Sejnowski [8] has been used by [13] to perform ICA. The algorithm was derived from the principle of optimal information transfer in neurons with sigmoidal transfer functions. It finds the matrix that represents the statistically independent vectors of the face images. Given that X is an input matrix, in which each row represent an image, then $U = WX$ is the output matrix of independent representation of the images, where W is an invertible weight matrix such that $X' = W^{-1}U$. The weight matrix, W , was found through an unsupervised learning algorithm that maximizes the mutual information between the input and the output of the nonlinear transformation [8]. The algorithm has proven successful for separating randomly mixed auditory signals and has been applied to natural scenes.

Regardless of which algorithm is used to compute ICA, there are two fundamentally different ways to apply ICA to face recognition. In Architecture I, the images are considered as random variables and the pixels as outcomes, while

Architecture II treaded the pixels as random variables and the images as outcomes.

A. Architecture I(ICA1): Statistically Independent Basis Images

In this architecture, the face images are variables and the pixel values provide observations for the variables. The images are organized as a data matrix $X^{m \times n}$, where m is the number of training images and n is the number of pixels. The input face images in X are considered to be a linear mixture of statistically independent basis images S combined by an unknown mixing

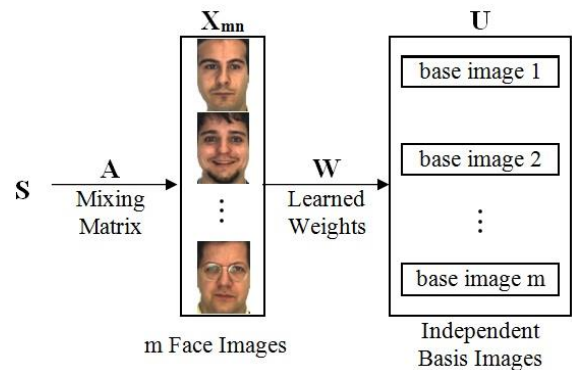


Fig. 1. Finding statistically independent basis images

matrix A . The weight matrix W is learnt by ICA algorithm such that the rows of $U=WX$ are as statistically independent as possible. The source images estimated by the rows of U are then used as basis images to represent faces (Fig. 1). The source separation, therefore, is performed in face space. Projecting the input images onto the learned weight vectors produces the independent basis images. The compressed representation of a face image is a vector of coefficients used for linearly combining the independent basis images to generate the image. Bartlett et al. [13] first apply PCA to project the data into a subspace of dimension m to control the number of independent components produced by ICA. The Information Maximization algorithm [8] is then applied to the eigenvectors to minimize the statistical dependence among the resulting basis images. Using PCA as preprocessing allows ICA to create subspaces of size m for any m . Liu and Wechsler [14] argue that applying ICA on the projected data enhances the performance by

- Discarding small trailing eigenvalues before whitening.
- Reducing the computational complexity by minimizing pair-wise dependences.

Let X be the matrix containing the zero-mean images and V be the matrix containing (in its columns) the first m eigenfaces that have the m highest eigenvalues. Then the PC representation of X is defined as $R_m = XV$. ICA is performed on V to produce the independent basis images matrix U , weight matrix W and sphering matrix W_z . The IC representation of the face images based on the set of m statistically independent basis images U is given by the rows of the matrix

$$F = R_m W_I^{-1}, \quad (1)$$

where $W_I = W W_z$ such that $W_I V^T = U$. The representation for test images is obtained by the following equation

$$F_{test} = R_{test} W_I^{-1}, \quad (2)$$

where $R_{test} = X_{test} V$.

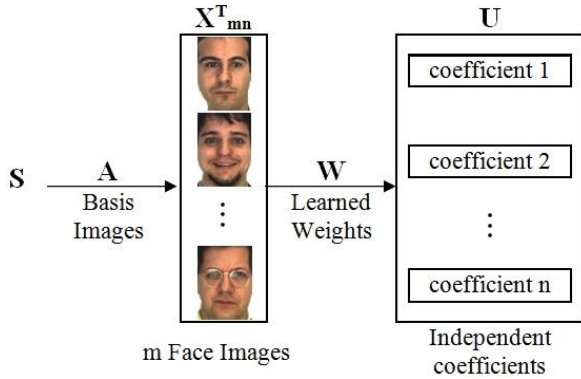


Fig. 2. Finding statistically independent coefficients

B. Architecture II(ICA2): A Factorial Face Code

In Architecture II the coefficients that represent the images are represented by ICA such that they are statistically independent. The data matrix X is organized so that rows represent different pixels and columns represent different images. The inverse matrix $A = W_I^{-1}$ contains the basis images in its columns. The statistically independent source coefficients (ICA representations) in S that comprise the input images are recovered in the columns of $U = W_I X$ (Fig. 2). Each column of U contains the coefficients of basis images in A for reconstructing each image in X such that each face $x = u_1 * a_1 + u_2 * a_2 + \dots + u_n * a_n$ where the column of U (u_1, u_2, \dots, u_n) is the ICA factorial representation. The representational code for test images is obtained by $W_I X_{test} = U_{test}$, where X_{test} is the zero-mean matrix of test images, and W_I is the weight matrix found by performing ICA on the training images.

C. Combined ICA Reconstruction System

A combined classifier "ICA1+2" was employed by [13] in which the summation of the similarity measures c_1 and c_2 of ICA1 and ICA2 respectively has been used as the similarity between test images and gallery (training) images. The similarity measure was evaluated by the cosine of the angle between the coefficient vector of the test image and the coefficient vectors of the training images

$$c = \frac{b_{test} \cdot b_{train}}{\|b_{test}\| \cdot \|b_{train}\|} \quad (3)$$

The cosines was used as a similarity measure because ICA performs significantly better than when using Euclidean distance [13]. The combined classifier was defined as $c_1 + c_2$, where c_1 and c_2 correspond to the similarity measure c in (3). According to [13] the classifier that combined the two ICA representations improved the performance of both ICA1 and ICA2 and outperformed PCA on all test sets. However, and according to extensive experiments, we found that many faces can be recognized using ICA1 while it cannot be recognized accurately using ICA2 and vice versa. Accordingly, we proposed a new classifier (Merging Distances) to utilize the advantages of both ICA architectures. It also utilizes the improvement of ICA1+2 in which the smallest similarity measure value resulted by ICA1+2 is integrated in the first rank of the new classifier.

III. THE PROPOSED CLASSIFIER (MERGING DISTANCES)

The proposed classifier was adopted in which the similarity measure vectors for both ICA1 and ICA2 were resorted in descending order and then integrated by merging the corresponding values of the two vectors to reconstruct a new similarity measure vector. The two similarity measure vectors v_{c1} and v_{c2} were combined as follows

$$v_c(i \times 2 - 1) = v_{c1}(i), \quad v_c(i \times 2) = v_{c2}(i) \quad (4)$$

where v_{c1} and v_{c2} is the similarity measure vectors for ICA1 and ICA2, respectively, $i = 2, \dots, n/2$, n is the number of ranks for both ICA1 and ICA2. By this way we ensure that the smallest similarity measure values of v_{c1} and v_{c2} will be in the early ranks of recognition which means that improvement in later ranks is guaranteed. In our experiments n has been chosen to be 30 ranks. To further utilizes the improvement of the combined classifier ICA1+2 [13], the smallest similarity measure value resulted by ICA1+2 is added in the first rank of the new similarity measure vector v_c . Consequently, the proposed classifier has the advantages of ICA1, ICA2, and ICA1+2. The combined classifier (Merging Distances) is deployed on v_c as a similarity measure.

IV. EXPERIMENTS AND DISCUSSION

This paper proposed a new classifier that combines the similarity measures of both ICA architectures ICA1 and ICA2 for face recognition across facial variation. We evaluated the performance of the proposed classifier compared with that of ICA1, ICA2, and ICA1+2. The AR Database has been used in the evaluation since it contains faces with different expression, illumination condition, and occlusions. 100 preprocessed subjects are used for training and testing. The first three images of each subject were used for training while 6 classes out of the 23 remaining images for testing. The 6 classes were chosen from the two sessions and contain expressions, illumination, and occlusions.

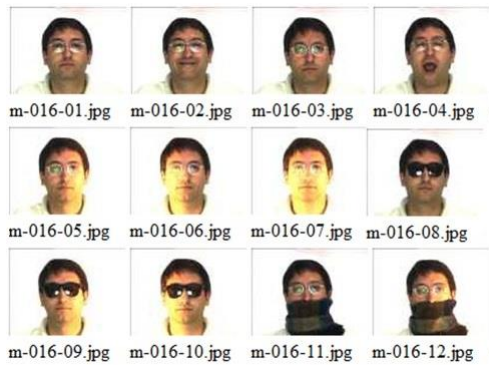


Fig. 3. Face Images for one subject from the AR face database: Examples of 12 expression recorded in the first session.

A. AR Face Database

A case study is conducted using AR Face database [15]. This database includes more than 4000 color images for 126 subject (70 men and 56 women). All images are frontal view faces with different facial expressions, illumination conditions, and occlusions (sun glasses and scarf). The pictures were taken at the CVC under strictly controlled conditions. Every subject has 26 face images recorded in two separate sessions. Fig. 4 shows 12 expressions out of 13 recorded in the first session for one individual from the AR face database. The data set was divided into training and testing groups, ICA was trained from the first session on first three images per subject, which represent neutral expression, smiling, and anger. While the algorithm is tested with 6 images per subject including "4: scream", "7: all side lights on", and "9: wearing sun glasses and left light on" from the first section, and "17: scream", "20: all side lights on", and "22: wearing sun glasses and left light on" from the second session.

B. Face recognition performance

The different Architectures of ICA (ICA1 and ICA2), and the combined classifier ICA1+2 were evaluated using AR

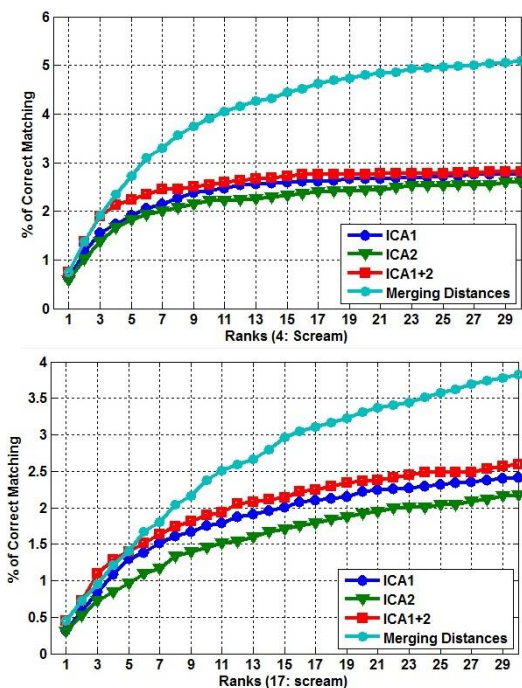


Fig. 4. Face recognition performance of (Merging Distances) compared to the individual classifiers ICA1 and ICA2 and the combined classifier ICA1+2 using images with scream expression

database. The recognition performance of the two architectures was evaluated by the nearest neighbor procedure using cosines as the similarity measure. The combined classifiers were deployed in which the similarity between a test image and a gallery image depends on both c_1 of ICA1 and c_2 of ICA2, where c_1 and c_1 correspond to the similarity measure. To analyze the performance of the two architectures of ICA algorithm we used Cumulative Mach Characteristic CMC.

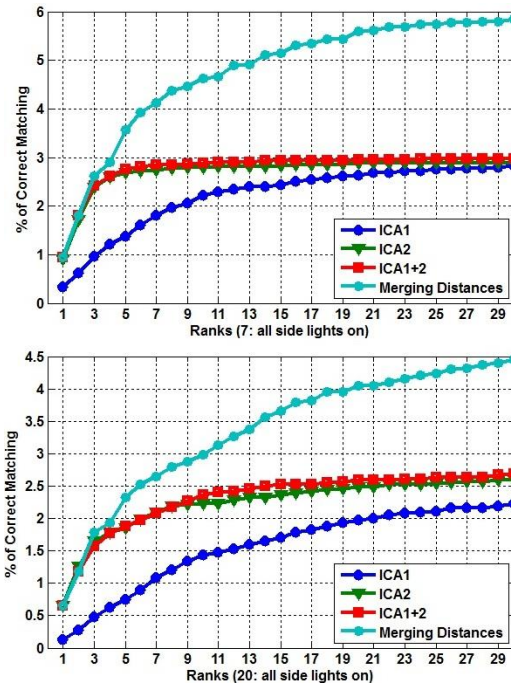


Fig. 5. Face recognition performance of (Merging Distances) compared to the individual classifiers ICA1 and ICA2 and the combined classifier ICA1+2 using images with illumination

1) Face images with scream expression and illumination: Fig. 4 and Fig. 5 show the Cumulative Mach Performance (CMC) of the performance of the ICA1, ICA2, ICA1+2, and the proposed classifier (Merging Distances) using face images with scream expression and illumination (e.g. all side lights on) taken in the same day and different day. As shown in the two figures, the proposed classifier consistently outperformed ICA1, ICA2, and ICA1+2 especially after rank 3.

2) Face images with occlusion: In case of face images with sun glasses, the proposed classifier has been tested using 2 testing sets including "wearing sun glasses and left light on" taken in the same day and different day. Fig. 6 shows the CMC using face images with sun glasses and left light on. The performance of the proposed classifier outperforms ICA1, ICA2, and ICA1+2 in later ranks. While in earlier ranks the performance of ICA1+2 is comparable with the proposed classifier. For the all testing faces taken in two different days,

it was noted that ICA1+2 performed better than both ICA1 and ICA2 in all ranks. However, it was comparable with "Merging Distances" in earlier ranks. By higher ranking, however, "Merging Distances" outperformed ICA1, ICA2 and ICA1+2. The overall results show that the proposed classifier has a consistence performance through different classes of the images in the database. While, the performance of ICA1 and ICA2 can be differ by changing the testing sets (See Fig. 4 and Fig. 5)

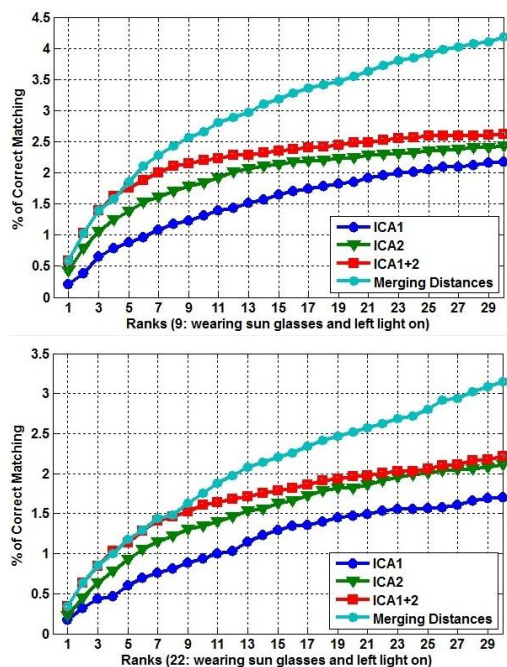


Fig. 6. Face recognition performance of the proposed classifier (Merging Distances) compared to the individual classifiers ICA1 and ICA2 and the combined classifier ICA1+2 using images with sun glasses and illumination

V. CONCLUSION

We have proposed a new classifier that merges the two ICA representations ICA1 and ICA2 for face recognition. The proposed classifier (Merging Distances) was employed in which the similarity measure vectors for both ICA representations were integrated by joining the corresponding values of the two vectors to reconstruct a new similarity measure vector. The two architectures of ICA were performed on face images in the AR Face Database. The new classifier that combined the two ICA architectures were tested using faces with occlusion, illumination and different expressions. Cumulative Match Characteristics was taken as a measure for evaluating the performance of the new classifiers. In the early ranking, the proposed classifier was comparable with ICA1,

ICA2 and the combined classifier ICA1+2 developed by [13], while it outperformed them in higher ranking (After rank 3). The new combined classifier achieves reasonable results in recognizing faces in all test cases. Future work will consider recognizing the face by using any non-occluded facial area.

REFERENCES

- [1] L. Xiaoguang, K. J. Anil, and C. Dirk, "Matching 2.5D Face Scans to 3D Models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, pp. 31-43, 2006.
- [2] X. Luan, B. Fang, L. Liu, W. Yang, and J. Qian, "Extracting sparse error of robust PCA for face recognition in the presence of varying illumination and occlusion," *Pattern Recognition*, vol. 47, pp. 495-508, 2// 2014.
- [3] I. Venkat, A. T. Khader, K. G. Subramanian, and P. De Wilde, "Recognizing occluded faces by exploiting psychophysically inspired similarity maps," *Pattern Recognition Letters*, vol. 34, pp. 903-911, 6/1/ 2013.
- [4] M. I. Razzak, M. K. Khan, and K. Alghathbar, "Bio-inspired Hybrid Face Recognition System for Small Sample Size and Large Dataset," in *Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2010 Sixth International Conference on*, 2010, pp. 384-388.
- [5] J. Lu and E. Zhang, "Gait recognition for human identification based on ICA and fuzzy SVM through multiple views fusion," *Pattern Recognition Letters*, vol. 28, pp. 2401-2411, 12/1/ 2007.
- [6] M. Girolami, *Advances in independent component analysis*. London [u.a.]: Springer, 2000.
- [7] T.-W. Lee, *Independent component analysis : theory and applications*. Boston: Kluwer Academic Publishers, 1998.
- [8] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural computation*, vol. 7, pp. 1129-1159, 1995.
- [9] A. J. Bell and T. J. Sejnowski, "The "independent components" of natural scenes are edge filters," *Vision research*, vol. 37, pp. 3327-3338, 1997.
- [10] S. Makeig, A. J. Bell, T.-P. Jung, and T. J. Sejnowski, "Independent component analysis of electroencephalographic data," *Advances in neural information processing systems*, pp. 145-151, 1996.
- [11] M. J. McKeown, S. Makeig, G. G. Brown, T.-P. Jung, S. S. Kindermann, A. J. Bell, et al. (1997). *Analysis of fMRI Data by Blind Separation into Independent Spatial Components*.
- [12] M. S. Bartlett and T. J. Sejnowski, "Viewpoint invariant face recognition using independent component analysis and attractor networks," *Advances in neural information processing systems*, pp. 817-823, 1997.
- [13] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski, "Face recognition by independent component analysis," *Neural Networks, IEEE Transactions on*, vol. 13, pp. 1450-1464, 2002.
- [14] C. Liu and H. Wechsler, "Comparative assessment of independent component analysis (ICA) for face recognition," in *International conference on audio and video based biometric person authentication*, 1999.
- [15] A. M. Martinez, "The AR face database," *CVC Technical Report*, vol. 24, 1998.

Modeling and Design of Anisotropic Circular Microstrip Patch Antenna Using Neurospectral Computation Approach

Sami BEDRA¹, Tarek FORTAKI², Randa BEDRA², Abderraouf MESSAI³

¹Industrial Engineering Department, University of Khenchela, 40004 Khenchela, Algeria

²Electronics Department, University of Batna, 05000 Batna, Algeria

³Electronics Department, University of Constantine1, 25000 Constantine, Algeria

bedra_sami@yahoo.fr

Abstract—in this paper, we propose a general design of circular microstrip antenna printed on isotropic or anisotropic substrate, based on artificial neural networks (ANN) in conjunction with spectral domain formulation. In the design procedure, synthesis ANN model is used as feed forward network to determine the resonant characteristics. Analysis ANN model is used as the reversed of the problem to calculate the antenna dimension for the given resonant frequency, dielectric constant, and height of substrate. The effective parameters were combined with artificial neural network in the analysis and the design of circular antenna to reduce the complexity of the spectral approach and to minimize the CPU time necessary to obtain the numerical results. The results obtained from the neural models are in very good agreement with the experimental results available in the literature. Finally, numerical results of the anisotropy substrate effect on the resonant characteristics are also presented.

Keywords—Circular Microstrip Antenna (CMSA), Artificial Neural Network (ANN), design and modeling, spectral domain approach, uniaxial anisotropy substrate.

I. INTRODUCTION

The microstrip antenna (MSA) is an excellent radiator for many applications such as mobile antenna, aircraft and ship antennas, remote sensing, missiles and satellite communications [1]. It consists of radiating elements (patches) photo etched on the dielectric substrate. Microstrip antennas are low profile conformal configurations. They are lightweight, simple and inexpensive, most suited for aerospace and mobile communication. Their low power handling capability posits these antennas better in low power transmission and receiving applications [2]. The flexibility of the Microstrip antenna to shape it in multiple ways, like square, rectangular, circular, elliptical, triangular shapes etc., is an added property. Various methods and commercial software are available for analysis and synthesis of microstrip antennas. These commercial design packages use computer intensive numerical methods such as, Finite Element Method (FEM), Method of Moment (MoM), Finite Difference Time Domain (FDTD) method, etc. These techniques require high computational resources and also take lot of computation time [3]. Even though all the losses can be directly included in the analysis, produced results may not provide satisfactory accuracy for all the cases. Because of these problems, Mishra and Patnaik have introduced the use of neural networks in conjunction with spectral domain approach to calculate the complex resonant frequency [4] and the input impedance [5] of rectangular microstrip resonators, this approach is named the neurospectral method. In reference [4],

the computational complexity involved in finding complex root is reduced, whereas, in reference [5], the neural network method evaluates the integrals appearing in the matrix impedance. Later on [6], Mishra and Patnaik have demonstrated the force of the neurospectral approach in patch antenna design by using the reverse modeling to determine the patch length for a given set of other parameters.

The increase in complexity of device modeling has led to rapid growth in the computational modeling research arena. To accommodate computational complexity, several computer aided design (CAD) modeling engines such as artificial neural networks (ANNs) were used [7-11]. ANNs, emulators of biological neural networks, have emerged as intelligent and powerful tools and have been widely used in signal processing, pattern recognition, and several other applications [9-10]. ANN is a massively parallel and distributed system traditionally used to solve problems of nonlinear computing [4, 12].

The objective of this work is to present an integrated approach based on artificial neural networks and electromagnetic knowledge (effective's parameters). We introduce the artificial neural networks in the analysis of circular antenna to reduce the complexity of the spectral approach and to minimize the CPU time necessary to obtain the numerical results. We have demonstrated the force of neurospectral approach in antenna modeling using ANN combined with EM knowledge to develop a neural network model for the calculation of resonant characteristics (resonant frequencies and bandwidths) of circular patch antenna printed

on isotropic or uniaxially anisotropic substrate. Using reverse modeling, ANN is built to find out the antenna dimensions for the given resonant frequency, dielectric constant and height of substrate. The models are simple, easy to apply, and very useful for antenna engineers to predict both patch dimensions and resonant characteristics of circular microstrip antenna taken into account the anisotropy in the substrate. To the best of our knowledge, this subject has not been reported in the open literature; the only published results on analysis of rectangular microstrip-patch resonators using neurospectral approach [4-6].

II. SPECTRAL DOMAIN FORMULATION

As seen in Fig. 1, the circular microstrip antenna (CMSA) consists of a patch of radius a , which is parallel to the ground plane; and this patch is separated from the ground plane by a dielectric substrate with relative permittivity ϵ_r , and thickness h . If we want to take the substrate uniaxial anisotropy's into account, the number of inputs increases; since the relative dielectric permittivity ϵ_r will be replaced by the pair of relative permittivities (ϵ_x, ϵ_z) , where ϵ_x and ϵ_z are the relative dielectric permittivity along x and z axis, respectively (Fig. 1).

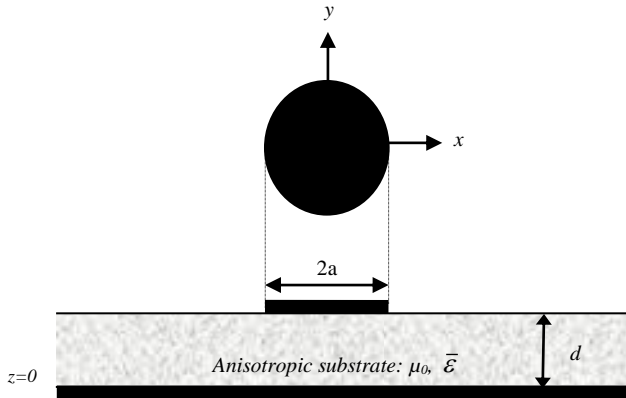


Fig. 1. Geometry of circular-disk microstrip antenna.

With the increase of design parameter's number, the network size increases, resulting in an increase in the size of training set required for proper generalization. Because of the different natures of the additional parameters, data generation becomes more complicated, a solution to this problem seems necessary. For the case of uniaxially anisotropic substrate, ϵ_{re} given in [13-14] there resulting values are:

$$\epsilon_{re} = \epsilon_z \quad (1)$$

$$d_e = d \sqrt{\frac{\epsilon_x}{\epsilon_z}} \quad (2)$$

In such an approach, the spectral function of Green, which binds the fields with the tangential electrical currents according to various plans of the drivers, must be given. Several techniques we proposed to evaluate the spectral Green function [14-15].

$$\mathbf{E}(\rho, \phi, z) = \begin{bmatrix} E_\rho(\rho, \phi, z) \\ E_\phi(\rho, \phi, z) \end{bmatrix} = \sum_{n=-\infty}^{n=+\infty} e^{in\phi} \int_0^\infty k_\rho dk_\rho \bar{\mathbf{H}}_n(\rho k_\rho) \cdot \mathbf{e}_n(k_\rho, z) \quad (3)$$

$$\mathbf{H}(\rho, \phi, z) = \begin{bmatrix} H_\phi(\rho, \phi, z) \\ -H_\rho(\rho, \phi, z) \end{bmatrix} = \sum_{n=-\infty}^{n=+\infty} e^{in\phi} \int_0^\infty k_\rho dk_\rho \bar{\mathbf{H}}_n(\rho k_\rho) \cdot \mathbf{h}_n(k_\rho, z) \quad (4)$$

$$\bar{\mathbf{H}}_n(\rho k_\rho) = \begin{bmatrix} J'_n(\rho k_\rho) & -\frac{in}{\rho k_\rho} J_n(\rho k_\rho) \\ \frac{in}{\rho k_\rho} J_n(\rho k_\rho) & J'_n(\rho k_\rho) \end{bmatrix} \quad (5)$$

In Eq. (5), $\bar{\mathbf{H}}_n(\rho k_\rho)$ is the kernel of the vector Hankel transform (VHT) [14-16], $J_n(\cdot)$ is the Bessel function of the first kind of order n , and the prime denotes differentiation with respect to the argument. The dagger implies conjugate transpose.

The relationship which relates the current on the conducting patch to the tangential electric field in the corresponding interface:

$$\mathbf{e}_n(k_\rho, z) = \bar{\mathbf{G}}(k_\rho) \cdot \mathbf{K}_n(k_\rho) \quad (6)$$

Where $\bar{\mathbf{G}}(k_\rho)$ dyadic Green's function in the vector Hankel transform domain [16]. Note that in the vector Hankel transform domain, the dyadic Green's function is diagonal and it is independent of the geometry of the radiating patch.

Note that, the tensor of Green for the considered structure can be easily determined. The tangential electric field is null on the conducting patch, which leads to an integral equation. To solve the integral equation, we apply the procedure of Galerkin which consists in developing the unknown distribution of the current on the circular patch is expanded into a series of basis functions [14-16]. The basis functions chosen in this article for approximating the current density on the circular patch are obtained from the model of the cavity. Boundary conditions require that the transverse components of the electric field vanish on the perfectly conducting disk and the current vanishes off the disk, to give the following set of vector dual integral equations:

$$\mathbf{E}_n(\rho, z) = \int_0^{+\infty} dk_\rho k_\rho \bar{\mathbf{H}}_n(k_\rho \rho) \cdot \bar{\mathbf{G}}(k_\rho) \cdot \mathbf{k}_n(k_\rho) = \mathbf{0}, \quad \rho < a \quad (7)$$

$$\mathbf{K}_n(\rho) = \int_0^{+\infty} dk_\rho k_\rho \bar{\mathbf{H}}_n(k_\rho \rho) \cdot \mathbf{k}_n(k_\rho) = \mathbf{0}, \quad \rho > a \quad (8)$$

The use of the method of the moments in the spectral

domain allows the resolution of the system of dual integral equations. The current on the disk is expressed in the form of a series of basis functions as follows:

$$\mathbf{K}_n(\rho) = \sum_{p=1}^P a_{np} \Psi_{np}(\rho) + \sum_{q=1}^Q b_{nq} \Phi_{nq}(\rho) \quad (9)$$

P and Q correspond to the number of basis functions of $\Psi_{np}(\rho)$ and $\Phi_{nq}(\rho)$, respectively, a_{np} and b_{nq} are the mode expansion coefficients to be sought. The corresponding VHT of the current is given by

Substitute the current expansion (10) into (7). Next, multiplying the resulting equation by $\rho \Psi_{nk}^+(\rho)$ ($k=1,2,\dots, P$) and by $\rho \Phi_{nl}^+(\rho)$ ($l=1,2,\dots,Q$), and while integrating from 0 to a , and using the Parseval's theorem for vector Hankel transform [16], we obtain a system of linear $P+Q$ algebraic equations for each mode n which can be written in the matrix form:

$$\bar{\mathbf{Z}}_n \cdot \mathbf{C}_n = \mathbf{0} \quad (11)$$

where:

$$\bar{\mathbf{Z}}_n = \begin{bmatrix} (\bar{\mathbf{Z}}_n^{\Psi\Psi})_{P \times P} & (\bar{\mathbf{Z}}_n^{\Psi\Phi})_{P \times Q} \\ (\bar{\mathbf{Z}}_n^{\Phi\Psi})_{Q \times P} & (\bar{\mathbf{Z}}_n^{\Phi\Phi})_{Q \times Q} \end{bmatrix}, \quad (12)$$

$$\mathbf{C}_n = \begin{bmatrix} (\mathbf{a}_n)_{P \times 1} \\ (\mathbf{b}_n)_{Q \times 1} \end{bmatrix}$$

Each element of the submatrices is given by:

$$\bar{\mathbf{Z}}_n^{vw}(i, j) = \int_0^{+\infty} dk \rho \mathbf{V}_{ni}^+(k, \rho) \cdot \bar{\mathbf{G}}(k, \rho) \cdot \mathbf{W}_{nj}(k, \rho) \quad (13)$$

where \mathbf{V} and \mathbf{W} represent either Ψ or Φ . For every value of the integer n , the system of linear equations (11) has non-trivial solutions when

$$\det[\bar{\mathbf{Z}}_n(\omega)] = 0 \quad (14)$$

This equation (14) is called characteristic equation of the structure (figure. 1). For the search of the complex roots of this equation, the method of Müller is used. It requires three initial guesses which must be close if possible to the sought solution to ensure a fast convergence.

Generally the real part (f_r) of the solution represents the resonant frequency of the structure, the imaginary part (f_i) indicates the losses of energy per radiation and the ratio ($2f_i/f_r$) gives the band-width (BW) and the quantities $Q=(f_r/2 f_i)$ stands for the quality factor [14-16].

In the following section, a basic artificial neural network is described briefly and the application of neural network to the prediction the resonant characteristics of the microstrip antenna are than explained.

III. NEURAL NETWORK MODELING

ANN learns relationships among sets of input-output data which are characteristic of the device under consideration. It is a very powerful approach for building complex and nonlinear relationship between a set of input and output data [17].

Artificial neural networks (ANNs) have been used frequently in signal processing applications, speech and pattern recognition, remote sensing, etc. for the last two decades [18]. Ability, adaptive capability and ease of implementation have made ANN a popular tool for many design problems in today's communication world [19]. More importantly, ANNs can generalize and respond correctly to slightly deviant input values, not presented during the training process [20]. These networks directly give good approximation to simulated and measured value, thereby avoiding the need for possibly a more complex and time-consuming conventional problem-specific algorithm [19]. In the present scenario, neural network models are used extensively for wireless communication engineering, which eliminates the complex and time-consuming mathematical and simulation procedures for designing antennas [21-23].

Multilayer perceptrons have been applied successfully to solve some difficult and diverse problems by training them in a supervised manner with a highly popular algorithm known as the error back propagation algorithm [23].

As shown in Fig. 2, the MLP consists of an input layer, one or more hidden layers, and an output layer. Neurons in the input layer only act as buffers for distributing the input signals x_i to neurons in the hidden layer. Each neuron in the hidden layer sums its input signals x_i after weighting them with the strengths of the respective connections w_{ji} from the input layer and computes its output y_j as a function f of the sum, namely

$$y_j = f\left(\sum w_{ji}x_i\right) \quad (15)$$

Where f can be a simple threshold function or a sigmoid or hyperbolic tangent function [24]. The output of neurons in the output layer is computed similarly.

Training of a network is accomplished through adjustment of the weights to give the desired response via the learning algorithms. An appropriate structure may still fail to give a better model unless the structure is trained by a suitable learning algorithm. A learning algorithm gives the change $\Delta w_{ji}(k)$ in the weight of a connection between neurons i and j at time k . The weights are then updated according to the formula

$$w_{ji}(k+1) = w_{ji}(k) + \Delta w_{ji}(k+1) \quad (16)$$

MLP can be trained using many different learning algorithms [25]. In this article, the following back propagation learning algorithm described briefly was used to train the MLP.

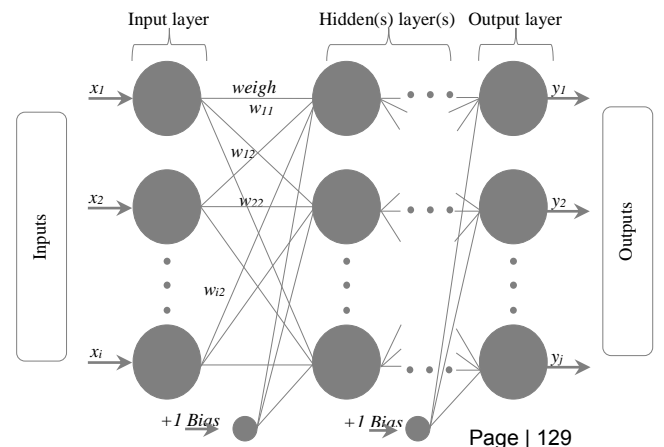


Fig. 2. General form of multilayered perceptrons.

The back-propagation algorithm is based on the error correction learning rule. Basically, error back propagation learning consists of two passes through the different layers of the network, a forward pass and a backward pass. In the forward pass, an activity pattern is applied to the sensory nodes of the network, and its effect propagates through the network layer by layer [25]. Finally, a set of outputs is produced as the actual response of the network. During the forward pass the synaptic weights of the networks are all fixed. During the backward pass, on the other hand, all the synaptic weights are adjusted in accordance with an error correction rule. The actual response of the network is subtracted from a desired response to produce an error signal. This error signal is then propagated backward through the network against the direction of synaptic connections. The synaptic weights are adjusted to make the actual response of the network move closer to the desired response in a statistical sense [23]. ANN models accuracy depends on the amount of data presented to it during training. A well-distributed, accurately simulated or measured and sufficient data are the basic requirement to obtain an efficient model. All the numerical results presented in this paper we obtained on a Pentium IV computer with a 2.6-GHz processor and a total RAM memory of 2 GB.

In this work, the patch geometry of the microstrip antenna is obtained as a function of input variables, which are height of the dielectric material (d_e), dielectric constants of the substrate (ϵ_{re}), and the resonant frequency (f_r), using ANN techniques “Fig. 3”. Similarly, in the analysis ANN, the resonant frequency of the antenna is obtained as a function of patch dimensions (a), height of the dielectric substrate (d_e), and dielectric constants of the material (ϵ_{re}) “Fig. 4”. Thus, the forward and reverse sides of the problem will be defined for the circular patch geometry in the following subsections.

Synthesis of the patch geometry of the microstrip antenna is a problem for which closed-form solutions exist. Therefore, this example is very useful for illustrating features and capabilities of synthesis ANN. Details of the problem are presented next.

A. The forward side of the problem: The synthesis ANN

The input quantities to the ANN black-box in synthesis “Fig. 3” can be ordered as:

- d_e : height of the dielectric substrate;
- ϵ_{re} : effective dielectric substrate;

- f_r : resonant frequency of the antenna.

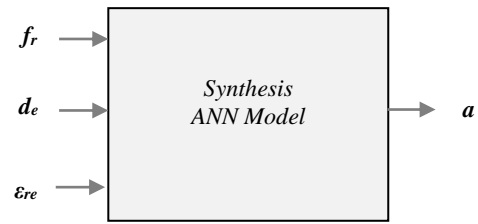


Fig. 3. Synthesis Neural model for predicting the patch geometry of circular microstrip antenna with effective parameters.

The following quantities can be obtained from the output of the black-box as functions of the input variables:

- a : radius of a circular patch;

B. The reverse side of the problem: The analysis ANN

In the analysis side of the problem, terminology similar to that in the synthesis mechanism is used, but the resonant frequency and the half-power bandwidth of the antenna are obtained from the output for a chosen dielectric substrate and patch dimensions at the input side as shown in “Fig. 4”

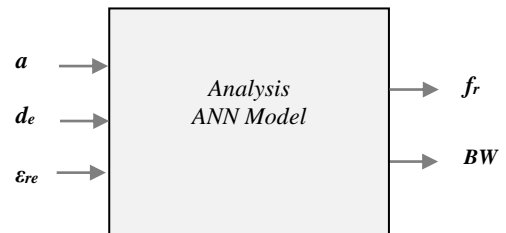


Fig. 4. Analysis Neural model for calculating the resonant frequency and half-power bandwidth of circular microstrip antenna with effective parameters.

The details of the network parameters for both these cases (analysis and synthesis) model are given in Table 1.

TABLE .1 COMPARISONS OF PERFORMANCE DETAILS OF ANALYSIS AND SYNTHESIS MODELS.

Algorithm details	Neurospectral approach	
	Analysis model	Synthesis model
Activation function	sigmoid	sigmoid
Training function (back-propagation)	trainrp	trainrp
Number of data	280	280
Number of neurons (input layer)	3	3
Number of neurons (hidden layers)	12-8	8-8
Number of neurons (output layer)	2	1
Epochs (number of iterations)	8000	8000
TPE (training performance error)	10^{-4}	10^{-4}
Time required	97 min	86 min
LR (learning rate)	0.6	0.5
MC (momentum constant)	0.7	0.6

IV. NUMERICAL RESULTS AND DISCUSSION

In order to confirm the computation accuracy of the neurospectral method, our results are compared with experimental and recent theoretical data [26-28]. Experimental and numerical evaluations have been performed with a patch for different radius a , printed on isotropic substrate ($\epsilon_x=\epsilon_z=2.43$) and thickness $d=0.49$ mm. The Table 2 summarizes our computed resonant frequencies and those obtained for TM₁₁ mode via spectral domain formulation [26-28]. The comparisons show a good agreement between our results and those of literature [26-28].

In the synthesis, neurospectral model give the best approximation to the target values. The results of the synthesis and comparison with targets are given in Table 3.

With the aim of confirming the computation accuracy for the case of uniaxially anisotropic substrate, we compare in Fig. 5 our results with theoretical data previously published [29].

TABLE 2. THEORETICAL AND EXPERIMENTAL VALUES OF THE RESONANT FREQUENCY FOR THE FUNDAMENTAL MODE OF CIRCULAR MICROSTRIP ANTENNA. $\epsilon_x = \epsilon_z = 2.43$, $d=0.49$ mm.

a (mm)	a/d	Experiment (GHz) [26]	Computed (GHz)			
			[26]	[27]	[28]	Present
1.9698	4.02	25.60	25.30	25.92	25.40	25.56
3.9592	8.08	13.10	13.30	13.55	13.30	13.18
5.8898	12.02	8.960	9.13	9.25	9.20	9.017
8.0017	16.33	6.810	6.80	6.87	---	6.823
9.9617	20.33	5.470	5.49	5.54	5.60	5.509

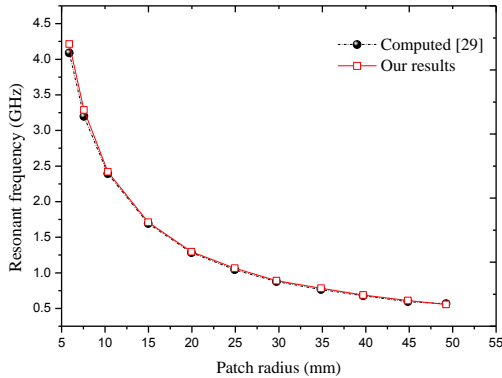


Fig. 5. Resonance frequency as a function of radius patch of a circular microstrip antenna on anisotropic substrate; Epsilam-10 ($\epsilon_x = 13$, $\epsilon_z = 10.3$), $d=1.27$ mm.

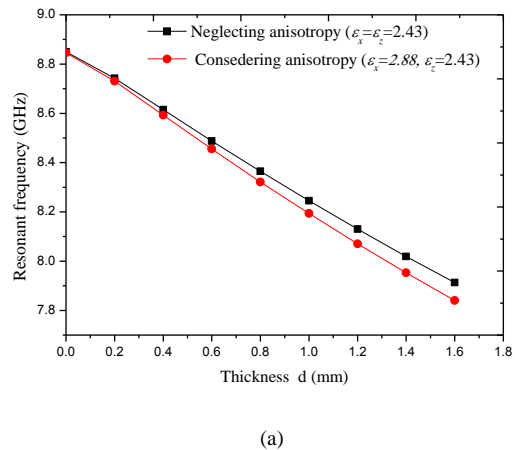
TABLE 3. REVERSE MODELING FOR THE PREDICTION OF ANTENNA DIMENSIONS.

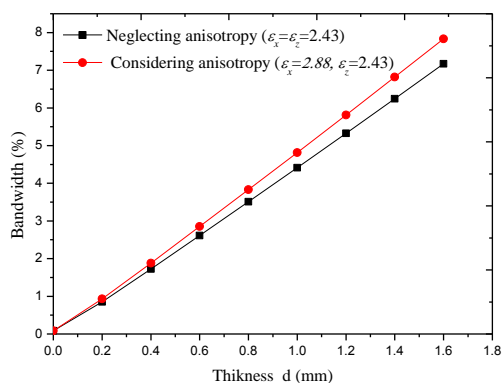
Input parameters			Target	ANN
d (mm)	$\epsilon_x = \epsilon_z$	f_r (GHz)	a (mm)	a (mm)
1.588	2.5	1.57	34.93	34.967
3.175	2.5	1.51	34.93	34.930
2.35	4.55	0.825	49.5	49.583
2.35	4.55	1.03	39.75	39.634
2.35	4.55	2.003	20	20.076
2.35	4.55	3.75	10.4	10.415
2.35	4.55	4.945	7.7	7.695
1.5875	2.65	4.425	11.5	11.565
1.5875	2.65	4.723	10.7	10.622
1.5875	2.65	5.224	9.6	9.596
1.5875	2.65	6.074	8.2	8.185
1.5875	2.65	6.634	7.4	7.402

It is seen from figure .5 that our results are close to those given in [29]. This validates the proposed model for the case of anisotropic substrate.

In Figure 6, results are presented for the resonant frequency and bandwidth of circular microstrip patch printed on an anisotropic dielectric substrate (PTFE).

In this figure, the results obtained for the resonant frequency and bandwidth of patch on anisotropic PTFE ($\epsilon_x=2.88$, $\epsilon_z=2.43$) are compared with the results that would be obtained if the anisotropy of Boron nitride were neglected ($\epsilon_x=\epsilon_z=2.43$). The patch has a radius of 6.35 mm.





(b)

Fig. 6. (a) Resonant frequency; (b) bandwidth of circular microstrip patch printed on anisotropic PTFE, the patch has a radius of 6.35mm.

The differences between the results obtained considering anisotropy and neglecting anisotropy reach 4.03 percent in the case of resonant frequencies and 32.34 percent in the case of half-power bandwidths. Thus, it can be concluded that the effect of uniaxial anisotropy on the resonant frequency and bandwidth of a circular microstrip patch antenna cannot be ignored and must be taken into account in the design stage.

V. CONCLUSION

A neural network-based CAD model can be developed for the analysis of a circular patch antenna printed on isotropic or anisotropic substrate, which is robust both from the angle of time of computation and accuracy. A distinct advantage of neuro-computing is that, after proper training, a neural network completely bypasses the repeated use of complex iterative processes for new cases presented to it. In the design procedure, syntheses ANN model is used as feed forward network to determine the resonant characteristics of circular microstrip antenna printed on anisotropic substrate. Analysis ANN model is used as the reversed of the problem to predict the antenna dimension for the given resonant frequency, dielectric constant and height of substrate. The spectral domain technique combined with the ANN method is several hundred times faster than the direct solution. This remarkable time gain makes the designing and training times negligible. Consequently, the Neurospectral method presented is a useful method that can be integrated into a CAD tool, for the analysis, design, and optimization of practical shielded (Monolithic microwave integrated circuit) MMIC devices.

REFERENCES

[1] G. Kumar, and K. P. Ray, "Broadband Microstrip Antennas" Artech House, London, 2003.
 [2] G. Garg, P. Bhartia, I. Bahl and A. Ittipiboon, "Microstrip Antenna Design Handbook," Artech House, Canton, 2001.

[3] P. P. Bhagat, D. Pujara, and D. Adhyaru, "Analysis and synthesis of microstrip patch antenna using Artificial Neural Networks," in Antennas and Propagation (APCAP), 2012 IEEE Asia-Pacific Conference on, 2012, pp. 120-121.
 [4] R.K. Mishra, A. Patnaik, "Designing rectangular patch antenna using the neurospectral method", IEEE Trans. Antennas and Propagat. Vol-51, Pp.1914 – 1921, Aug 2003.
 [5] R.K. Mishra, A. Patnaik, "Neurospectral computation for complex resonant frequency of microstrip resonators", IEEE Microwave and Guided Wave Letters, VOL. 9, NO. 9, pp.351-353, SEP 1999.
 [6] R.K. Mishra, A. Patnaik, "Neurospectral computation for input impedance of rectangular microstrip antenna", Electron. Lett., Vol-35, pp.1691 – 1693, 30 Sep 1999.
 [7] Y. Tighilt, F. Bouttout, and A. Khellaf, "Modeling and design of printed antennas Using neural networks", Int J RF and Microwave CAE., Vol. 21, pp.228–233, 2011.
 [8] V. T. Vandana, Pramod Singhal, "Microstrip Antenna Design Using Artificial Neural Networks" Int J RF and Microwave CAE 20: 76–86, 2010.
 [9] D. Vijay, Lakshman M Srinivas V, Vani Ch Yuriy G, Tayfun O, "Sensitivity Driven Artificial Neural Network Correction Models for RF/Microwave Devices" Int J RF and Microwave CAE., Vol. 22, pp.30–40, 2012.
 [10] Siakavara K., "Artificial neural network based design of a three-layered microstrip circular ring antenna with specified multi-frequency operation". Neural Comput & Applic., Vol.18, pp.57–64, 2009.
 [11] F. Wang, V.K. Devabhaktuni, and Q.J. Zhang, "Neural network structures and training algorithms for RF and micro-wave applications", Int J RF Microwave Comput Aided Eng., Vol. 9, pp. 216–240, 1999.
 [12] S. Kulshrestha, Deven J. Chheda, S.B. Chakrabarty, Rajeev Jyoti & S.B. Sharma, "Pole discontinuity removal using artificial neural networks for microstrip antenna design," International Journal of Electronics, Vol. 98:12, 1711-1720, 2011.
 [13] F. Bouttout, F. Benabdelaziz, A. Benghalia, D. Khedrouche, and T. Fortaki, "Uniaxially anisotropic substrate effects on resonance of rectangular microstrip patch antenna," Electron. Lett., Vol. 35, No. 4, 255-256, 1999.
 [14] T. Fortaki, D. Khedrouche, F. Bouttout and A. Benghalia, "Vector Hankel transform analysis of a tunable circular microstrip patch", Commun. Numer. Meth. Engng, 21:219-231, 2005.
 [15] W. C. Chew, T. M. Habashy, "The use of vector transforms in solving some electromagnetic scattering problems". IEEE Trans. Antennas and Propagat., (7):871–879, 1986.
 [16] S. Bedra, R. Bedra, S. Benkouda, and T. Fortaki, "Full-wave analysis of anisotropic circular microstrip antenna with air gap layer," Progress In Electromagnetics Research M, vol. 34, pp. 143-151, 2014.
 [17] K. Guney, C. Yildiz, S. Kaya, and M. Turkmen, "Artificial neural networks for calculating the characteristic impedance of air-suspended trapezoidal and rectangular-shaped microshield lines," Journal of Electromagnetic Waves and Applications, vol. 20, pp. 1161-1174, 2006.
 [18] P. Chopra and M. Chandrasekhar, "ANN modeling for design of a matched low noise pHEMT amplifier for mobile application," Journal of Computational Electronics, pp. 1-9, 2013.
 [19] T. Bose and N. Gupta, "Design of an aperture-coupled microstrip antenna using a hybrid neural network," Microwaves, Antennas & Propagation, IET, vol. 6, pp. 470-474, 2012.
 [20] S. Kulshrestha, D. J. Chheda, S. Chakrabarty, R. Jyoti, and S. Sharma, "Pole discontinuity removal using artificial neural networks for microstrip antenna design," International Journal of Electronics, vol. 98, pp. 1711-1720, 2011.
 [21] R. K. Mishra and A. Patnaik, "Neural network-based CAD model for the design of square-patch antennas," Antennas and Propagation, IEEE Transactions on, vol. 46, pp. 1890-1891, 1998.
 [22] A. Patnaik and R. K. Mishra, "ANN techniques in microwave engineering," Microwave Magazine, IEEE, vol. 1, pp. 55-60, 2000.
 [23] K. Kumar and N. Gunasekaran, "Bandwidth enhancement of a notch square shaped microstrip patch antenna using neural network approach," in Emerging Trends in Electrical and Computer Technology (ICETECT), 2011 International Conference on, 2011, pp. 797-799.

- [24] K. Guney and S. Gultekin, "A comparative study of neural networks for input resistance computation of electrically thin and thick rectangular microstrip antennas," *Journal of Communications Technology and Electronics*, vol. 52, pp. 483-492, 2007.
- [25] S. Haykin, *Neural networks: a comprehensive foundation*: Prentice Hall PTR, 1994.
- [26] V. Losada, R. R. Boix, and M. Horn, "Resonant modes of circular microstrip patches in multilayered substrate," *IEEE Trans. Antennas propagat.*, Vol. 47, No. 4, pp.488-497, 1999.
- [27] F. Benmeddour, C. Dumond, F. Benabdelaziz, and F. Bouttout, "Improving the performances of a high Tc superconducting circular microstrip antenna with multilayered configuration and anisotropic" *Progress In Electromagnetics Research C*, Vol. 18, 169-183, 2011.
- [28] A. Motevasselian, "Spectral domain analysis of resonant characteristics and radiation patterns of a circular disc and an annular ring microstrip antenna on uniaxial substrate" *Progress In Electromagnetics Research M*, Vol. 21, 237-251, 2011.
- [29] A. K. Verma, and Nasimuddin, "Analysis of circular microstrip patch antenna as an equivalent rectangular microstrip patch antenna on iso/anisotropic thick substrate," *IEE proc-Microw. Antennas propag.*, vol. 150, No. 4, 2003.

Bridging the Gap between Modeling of Mobile Agent-based Systems and Semantic Web using Meta-Modeling and Graph Grammars

Aissam Belghiat^{1,2}

¹Département d'informatique, Université 20 Août 1955-Skikda
Skikda 21000, Algérie
belghiatissam@gmail.com

*Allaoua Chaoui*²

²MISC Laboratory, Department of Computer Science, University of Constantine2
Constantine 25000, Algeria
a_chaoui2001@yahoo.com

Ali Aldahoud

Al-Zaytoonah University of Jordan,
P.O. Box 130, Amman 11733, Jordan
aldahoud@zuj.edu.jo

Abstract—Recently, the mobile agent-based paradigm has received more attention especially in distributed systems where it provides multiple solutions to several problems which can't be resolved by the object-based paradigm. Unfortunately, this paradigm lacks the interconnection with semantic web standards such as the language OWL (Ontology Web Language) which make it far from profiting from research results and advances in this area. We try in this paper to bridge the gap between the two domains by proposing an integrated approach for modeling mobile agent-based software systems using a transformation of mobile class diagrams into OWL ontologies. The developed approach allows interconnection of mobile agent and Semantic Web technologies can be used in a mobile agent-based application where such interconnection is needed. We use the meta-modeling and graph grammars tool AToM³.

Keywords—M-UML; OWL; MDA; Graph Transformation; AToM³

I. INTRODUCTION

Mobile agent-based software systems are increasingly very complex; actually the development of such systems is a difficult task since the great number of constraints that are evolved during the development process such as the mobility and security. Modeling and designing of mobile agent-based software systems have received important attention in the last years to deal with previous problems [18]. M-UML [19] has been introduced as an extension of UML [7] for modeling mobile agent-based systems [26]. Researchers have tried to relate mobile agent paradigm to the object oriented paradigm using the standard of object oriented modeling to model mobile agents by introducing to it the appropriate artifacts in order to support the new paradigm.

In other side, the ontologies provide explicit and formal specifications of shared conceptualizations; they are described formally using description logics implemented in OWL language. The knowledge generated from an M-UML diagram

during the software development process is a valuable asset in particular in the analysis and design tasks. In order to profit from it, they must be represented and stored in ontologies and will be used for reasoning on mobile agent based software systems.

In this paper, which extends our previous work [20], we propose a set of rules for transforming mobile class diagrams into ontologies described in OWL language in order to profit from the power of ontologies. So, the knowledge described by those diagrams can be reused, shared and linked with other information. The meta-modeling tool AToM³ is used to create meta-models for mobile class diagram and OWL models. A graph grammar is proposed for automatic transformation between the two formalisms.

The rest of the paper is organized as follows. In Section 2, we present some related works. In Section 3, we present some basic notions about M-UML, OWL, and graph grammars. In Section 4, we describe our approach of transforming M-UML

class diagrams to OWL ontologies. In Section 5, we illustrate our approach through an example. Finally concluding remarks and perspectives are presented in Section 6.

II. RELATED WORKS

Several works exist in the literature in the context of extracting ontologies from UML diagrams. In [14] the authors have proposed a transformation of UML towards DAML by showing similarities and differences between the two parts of the translation. In [15] the authors have proposed a transformation of a profile UML OUP (Ontology UML Profile) towards an ontology OWL. In [6], the OMG remarks the interest of such subject, it then immediately proposed the ODM which provides a profile for writing RDF and OWL within UML. The ODM also includes partial mappings between UML and OWL. It should be noted that several works are performed as answer to the call of the OMG and gathered in the ODM and it is impossible to evoke all of them here. In [9], an implementation of the ODM using ATL language is presented. In [5], the author has applied a style sheet on a XMI file which represents a model of a class diagram to generate an ontology OWL DL represented as RDF/XML format. In [16], a detailed comparison between UML and OWL has been developed.

On the other hand, AToM³ has been adopted to be a very powerful tool which the meta-modeling and the transformations between formalisms. In fact, in [21] the authors have proposed a formal framework and a tool for the specification and verification of G-Nets models using graph transformation. In [20] the authors have developed an AToM³ based approach for the automatic generation of OWL ontologies from UML diagrams. In [23], the authors have developed an approach for modeling and analysis of mobile agent-based software systems by transforming M-UML statecharts models to nested nets models. In [22], an approach for transforming mobile activity diagrams to nested Petri nets models has been proposed. Also in [1, 2, 3, 17, 18, and 25] we can find treatment and translation of multiple UML diagrams. In these works the Meta modeling allows visual modeling and graph grammar allows the transformation among them.

In contrast to all these previous works, we have the first who think to relate the modeling of mobile agent-based systems and semantic web by translating a profile of UML class diagram for mobile systems into the OWL.

III. BACKGROUND

The main contribution of the paper is to develop an integrated environment based on meta-modeling and graph grammars for modeling and analysis of mobile agent based software systems which are modeled as a set of mobile class diagrams. We recall here some notions about M-UML and OWL.

A. Mobile Class Diagram

A M-UML Class Diagram [19] has been introduced to describe the static structure of a mobile system by showing all relationships between different types of classes. Mobile objects/agents are created by the instantiation of a mobile class shown with a box (M). A class inherits the mobility by the

relation of inheritance while not necessarily true by aggregation relation. An affected class shown with a dashed box (M) is a class that contains methods (behavior) which communicate with other mobile classes. A remote class shown with a dashed box (R) is a class that contains methods (behavior) which communicate with a remote mobile object/agent. A mobile object/agent invokes methods which will be labeled depending on the location of it with either (M) or (R) and a class that includes the two types will show with dashed boxes (M) and (R). Figure 1 [19] shows an example of a mobile class diagram.

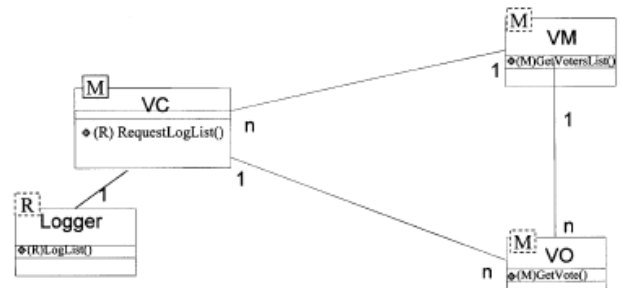


Fig. 1. A mobile class diagram.

B. OWL

OWL (Ontology Web Language) is a language for representing ontologies by defining the concepts of a domain and the relations between them what will allow automatic reasoning about the domain knowledge using their formal semantics. OWL1 offers three sublanguages with increasing expression oriented for specific communities of developers and users: OWL Lite, OWL DL, and OWL Full [10] whereas OWL2 defines three new profiles: OWL2 EL, OWL2 QL, and OWL2 RL [13].

C. Graph Grammars

Graph transformation was largely used for the expression of model transformation [4]; particularly, transformations of visual models can be naturally formulated by graph transformation, since the graphs are well adapted to describe the fundamental structures of models [17]. The set of graph transformation rules constitutes what is called the model of graph grammar. A graph grammar is a generalization, for graphs, of Chomsky grammars. Each rule of a graph grammar is composed of a left hand side (LHS) pattern and of a right-hand sided (RHS) pattern. Therefore, the graph transformation is the process to choose a rule among the graph grammar rules, apply this rule on a graph pattern that matches the LHS pattern to produce the RHS pattern, and reiterate the process until no rule can be applied [4]. We have adopted the AToM³ tool [1] which is a visual tool for model transformation to implements our approach. In the next sections, we will discuss how we use AToM³ to meta-model mobile class diagrams and how to generate OWL models by applying a graph grammar.

IV. OUR APPROACH

A. Overview

Mobile agent-based software systems are very difficult to design and to implement, although of this, we are urgently and still need this type of software systems to resolve some huge problems in different domains that the object oriented paradigm could not deal with them. We propose in this paper integrated approach M-UML class diagram/OWL ontology for modeling and analysis of mobile agent-based software systems by direct transformation of the mobile class diagram to OWL ontology. A mobile class diagram describes statically and in a very rich way the entities evolved in a mobile agent-based software system and all relationships among them. This ontology generated contains the knowledge extracted from the M-UML diagram and will be used for reuse purpose, knowledge sharing, conversation, integration and reasoning on mobile agent-based software systems.

The architecture of the proposed approach (Figure 2) is based on meta-modeling and graph grammars. For the realization of this application, we have to propose a meta-model of mobile class diagrams at AToM³ canvas. In addition, we have to develop a graph grammar made up of several rules which allows transforming progressively all what is modeled on the canvas towards an OWL ontology in RDF/XML format stored in a disk file. The graph grammar is based on transformation rules; each rule deals with some constructs in the left hand side (LHS) to transform them to others constructs in the right hand side (RHS). For ontology, the choice among OWL profiles is made on OWL DL because it places certain constraints on the use of the structures of OWL such as separation between classes, data types, data type properties, object properties, annotation properties, ontologies properties, individuals, data values, and integrated vocabulary [11][12].

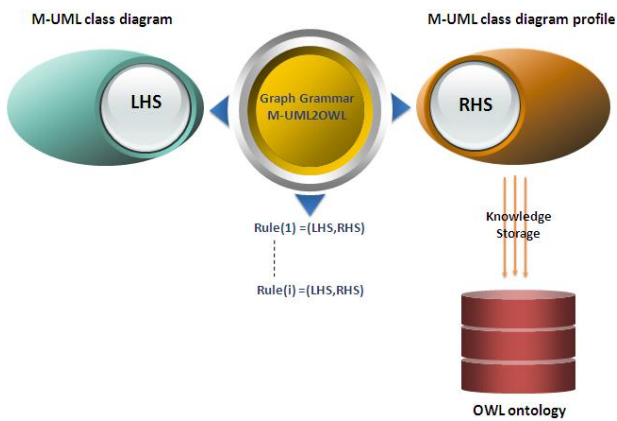


Fig. 2. Architecture of the proposed approach.

B. Transformation rules

We propose a set of rules to transform classes, enumerations, associations, roles, dependencies, association classes, and all the elements of a mobile class diagram that are important to store in the OWL ontology. For lack of space, we

have presented class (and their extensions) transformation rules in table 1.

Concerning the transformation of data types, all data types used in M-UML are transformed into XML schema (XSD) data types because OWL uses the majority of the datatypes integrated into XML schema. The calls of these data types are done through data type URI address reference <http://www.w3.org/2001/XMLSchema>[11]. The instances of the primitive types used in M-UML itself include: Boolean, Integer, String, Unlimited Natural[7].

TABLE I. TRANSFORMATION RULES.

M-UML to OWL	
Remote Class	
	<pre> <owl:DatatypePropertyrdf:ID="isRemote"> <rdfs:domainrdf:resource="#Remote-ClassName"/> <rdfs:rangerdf:resource="http://www.w3.org/2001/XMLSchema#boolean"/> </owl:DatatypeProperty> <owl:Classrdf:ID="Remote-ClassName"> <rdfs:subClassOf> <owl:Restriction> <owl:onPropertyrdf:resource="isRemote"/> <owl:mincardinalityrdf:datatype="http://www.w3.org/2001/XMLSchema#nonNegativeInteger">1 </owl:Restriction> </rdfs:subClassOf> </owl:Class>...</pre>
<p>A remote class is represented by a data type property and a class OWL. A remote class can contain behaviors affected by multiple remote mobile agent, we use the restriction <i>mincardinality</i> to indicate this.</p>	
Mobile Class	
	<pre> <owl:DatatypePropertyrdf:ID="isMobile"> <rdfs:domainrdf:resource="#Mobile-ClassName"/> <rdfs:rangerdf:resource="http://www.w3.org/2001/XMLSchema#boolean"/> </owl:DatatypeProperty> <owl:Classrdf:ID="Mobile-ClassName"> <rdfs:subClassOf> <owl:Restriction> <owl:onPropertyrdf:resource="isMobile"/> <owl:cardinalityrdf:datatype="http://www.w3.org/2001/XMLSchema#nonNegativeInteger">1 </owl:Restriction> </rdfs:subClassOf> </owl:Class>...</pre>
<p>A mobile class is represented by a data type property and a class OWL. A mobile class can contain exactly one property concerning the mobility of their objects.</p>	
Affected Class	

```

    <owl:DatatypeProperty rdf:ID="isAffected">
    <rdfs:domain rdfs:resource="#Affected-ClassName"/>
    <rdfs:range rdfs:resource="http://www.w3.org/2001/XMLSchema#boolean"/>
    </owl:DatatypeProperty>

    <owl:Class rdf:ID="Mobile-ClassName">
    <rdfs:subClassOf>
    <owl:Restriction>
    <owl:onProperty rdf:resource="#isAffected"/>
    <owl:minCardinality rdf:datatype="http://www.w3.org/2001/XMLSchema#nonNegativeInteger">1
    </owl:minCardinality>
    </owl:Restriction>
    </rdfs:subClassOf>
    </owl:Class>...
    
```

An affected class is represented by a data type property and a class OWL. An affected class can contain behavior (methods) that is affected (communicating with) by mobile objects, we use the restriction *minCardinality* to indicate this.

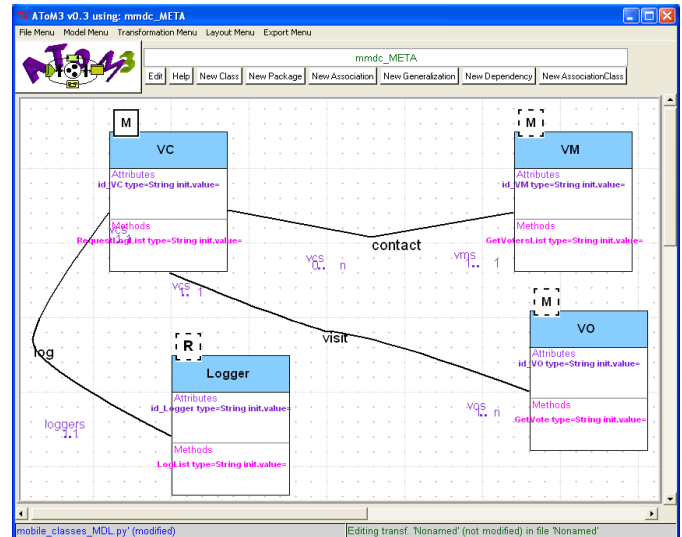


Fig. 4. Generated tool for mobile class diagram.

C. Meta-model of UML Class diagram

To build M-UML class diagrams in ATOM³, we have to define a meta-model for them. Our meta-model is composed of two classes and four associations developed by the meta-formalism (CD_classDiagramsV3), and the constraints are expressed in Python [8] code (Figure 3).

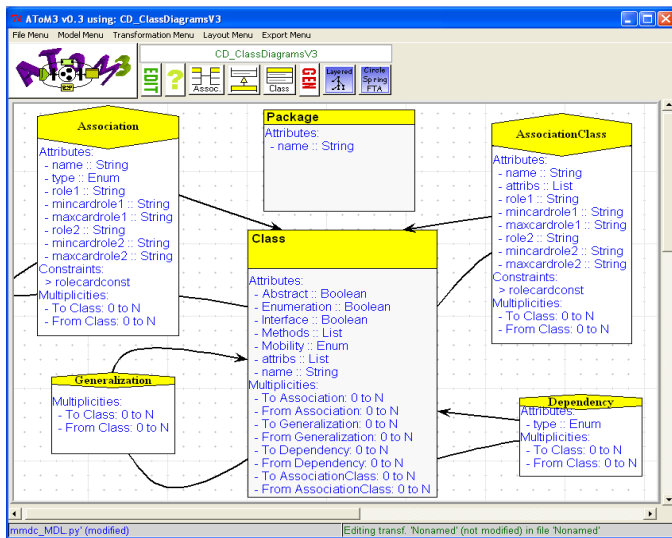


Fig. 3. Mobile class diagram meta-model

After building our meta-model, it remains only its generation. The generated meta-model comprises the set of classes modeled in the form of buttons which are ready to be employed for a possible modeling of a class diagram. Figure 4 shows an example of a mobile class diagram of a mobile voting system [19] modeled in our mobile class diagram environment.

D. The Proposed Graph grammar

To perform the transformation between class diagrams and OWL ontologies, we have proposed a graph grammar composed of an initial action, ten rules, and a final action. For lack of space, and because we used python code to specify the transformation in the condition and action of each rule, we have not presented all the rules.

Initial Action: Ontology header

Role: In the initial action of the graph grammar, we create a file with sequential access in order to store generated OWL code. To do that, we used Python. We begin by writing the ontology header which is fixed for all generated ontologies (Figure 5).

```

    self.rewritingSystem.inc = ATOM3String(1,20)
    obFichier = open('owlcode.owl','w')
    obFichier.write('<?xml version="1.0" encoding="UTF-8"?> \n')
    obFichier.write('<rdf:RDF \n')
    obFichier.write('  <xml:base="http://MobileUML.to/owl#" \n')
    obFichier.write('  <xmlns:xsd="http://www.w3.org/2001/XMLSchema#" \n')
    obFichier.write('  <xmlns:owl="http://www.w3.org/2002/07/owl#" \n')
    obFichier.write('  <xmlns:rdfs="http://www.w3.org/1999/02/22-rdf-syntax-ns#" \n')
    obFichier.write('  <xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#" \n')
    obFichier.write('  <xmlns="http://MobileUML.to/owl#" \n')
    obFichier.write('<owl:Ontology rdf:about="#Mobile_CD_Ontology"> \n')
    obFichier.write('<rdf:label>Written by Aissam BELGHIT</rdf:label> \n')
    obFichier.write('</owl:Ontology> \n')
    obFichier.close()
    
```

Fig. 5. Ontology header definition.

Final Action: the end of ontology

Role: In the final action of the graph grammar, we end our ontology. So, we will have to open our file and to add '</rdf:RDF>' (Figure 6).

the OWL file to add the adequate OWL code of this class in the action of the rule.

TABLE II. TRANSFORMATION OF DIFFERENT TYPES OF M-UML CLASSES.

V. EXAMPLE

Let us apply our approach on the mobile class diagram illustrated in figure 4 which is borrowed from [19]. It models a mobile voting system, where a mobile agent VC (vote collector) gets a list of voters from a stationary agent VM (vote manager) and visits the VO's (voters) stations those already have the list of candidates to collect votes and return them to the VM that mandated the VC in action. It should be noted that this example does not claim to be exhaustive but it gathers most important elements of a mobile class diagram such as: mobile class, affected class, remote class, association, attributes and different types of methods. By applying our graph grammar on this example, we have first obtained the intermediate graphs shown in figure 7.

Condition

```
node = self.getMatched(graphID, self.LHS.nodeWithLabel(1))
return not hasattr(node, "rule executed")
```

LHS

=

RHS

Action

```
node = self.getMatched(graphID, self.LHS.nodeWithLabel(1))
classname = node.name.getValue()
node.rule_executed = True
mob = node.Mobility.getValue()[1]
abst = node.Abstract.getValue()[1]
interf = node.Interface.getValue()[1]
if abst == 1:
    self.getMatched(graphID,
self.LHS.nodeWithLabel(1)).name.setValue('Abstract-'+classname)
elif interf == 1:
    self.getMatched(graphID,
self.LHS.nodeWithLabel(1)).name.setValue('Interface-'+classname)
obFichier = open('owlcode.owl', 'a')
node = self.getMatched(graphID, self.LHS.nodeWithLabel(1))
classname = node.name.getValue()
if mob == 1:
    obFichier.write('<owl:Class
rdf:ID="' + "Mobile-" + classname + "' + "/> \n')
    self.getMatched(graphID,
self.LHS.nodeWithLabel(1)).name.setValue('Mobile-'+classname)
elif mob == 2:
    obFichier.write('<owl:Class
rdf:ID="' + "Affected-" + classname + "' + "/> \n')
    self.getMatched(graphID,
self.LHS.nodeWithLabel(1)).name.setValue('Affected-'+classname)
elif mob == 3:
    obFichier.write('<owl:Class
rdf:ID="' + "Remote-" + classname + "' + "/> \n')
    self.getMatched(graphID,
self.LHS.nodeWithLabel(1)).name.setValue('Remote-'+classname)
```

Editing ATOM3constraint

```
obFichier = open('owlcode.owl', 'a')
obFichier.write('\n' + '<rdf:RDF>')
obFichier.close()
```

Fig. 6. End of ontology.

Rule 1: Class transformation

Name: class2class

Priority: 1

Role: This rule transforms an M-UML class (all type of classes) towards an OWL class (cf. Table 2). In the condition of the rule we test if the class is already transformed. If not yet, we reopen

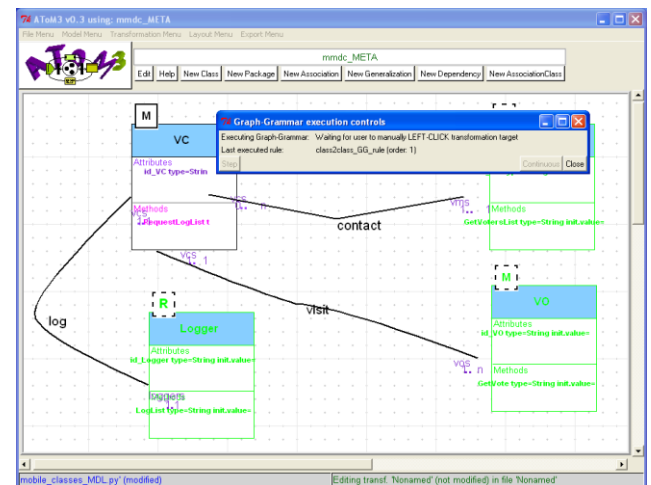


Fig. 7. Intermediate graph

Then we have obtained the graph of figure 8 after the termination of execution of the graph grammar.

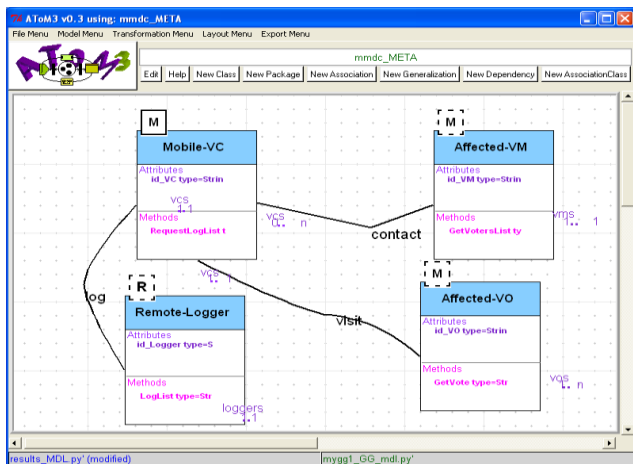


Fig. 8. Class diagram after execution

In parallel, there is an automatic generation of the file containing OWL code stored on hard disc validated and visualized using SWOOP [24] (Figure 9, 10, 11):

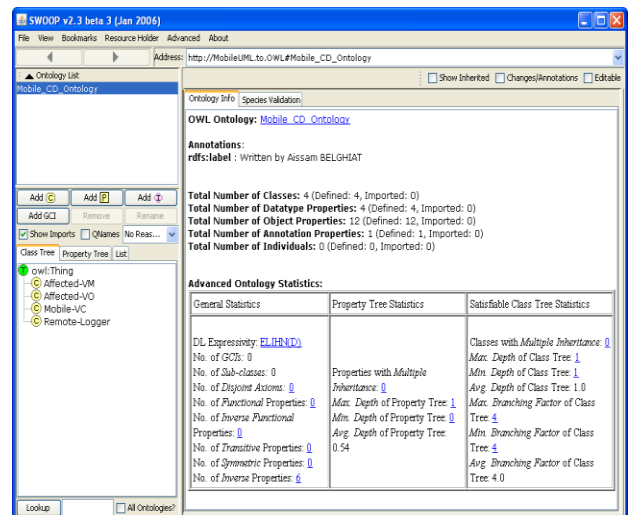


Fig. 10. The OWL ontology properties.

VI. CONCLUSION

The objective of this work is to develop an integrated environment for modeling and analysis of mobile based software systems by the transformation of M-UML class diagrams to OWL ontologies. The approach has been implemented using the ATOM³ tool. For the realization of this application we have developed a meta-model for M-UML class diagrams, and a graph grammar named “M-UML2OWL” composed of several rules which enables us to transform a mobile class diagram to an OWL ontology stored in a hard disk file. The generated ontology will be used for reuse purpose, knowledge sharing, conversation, integration and reasoning on mobile agent-based software systems.

In future work, we plan to generalize the extraction of OWL ontologies from others M-UML diagrams since they represent different aspects of the systems. We plan also to realize this transformation using other graph transformation tools such as Triple Graph Grammars [27] which provide bidirectional transformations.

REFERENCES

- [1] ATOM³. Home page: <http://atom3.cs.mcgill.ca.2002>.
- [2] J. D. Lara, H. Vangheluwe, “Computer aided multi-paradigm modeling to process petri-nets and statecharts,” International Conference on Graph Transformations (ICGT), Lecture Notes in Computer Science, vol. 2505, pp. 239-253, Springer-Verlag, Barcelona, Spain, 2002.
- [3] J. D. Lara, H. Vangheluwe, “Meta-modeling and graph grammars for multi-paradigm modeling in ATOM³,” Software and Systems Modeling, Vol. 3, pp. 194-209, Springer-Verlag, Special Section on Graph Transformations and Visual Modeling Techniques, 2004.
- [4] G. Karsai, A. Agrawal, “Graph Transformations in OMG’s Model-Driven Architecture,” Lecture Notes in Computer Science, Vol 3062, 243-259, Springer Berlin /Heidelberg, juillet 2004.
- [5] Sebastian Leinhos, <http://diplom.ooyoo.de>, 2006.
- [6] OMG, “Ontology Definition Metamodel”, V1.0, <http://www.omg.org/spec/ODM/1.0>, May 2009.
- [7] OMG, “OMG Unified Modeling Language, Infrastructure, v2.3”, <http://www.omg.org/spec/UML/2.1.2/Infrastructure/> PDF, May 2010.
- [8] Python. Home page: <http://www.python.org>.

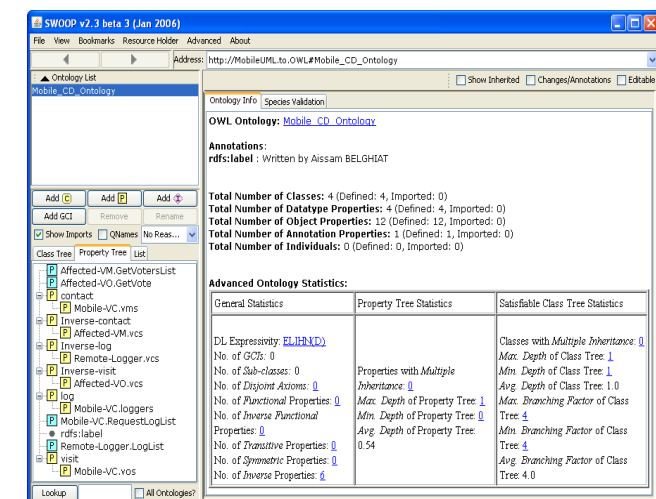


Fig. 9. The OWL ontology classes.

- [9] SIDo Group, “ATL Use Case - ODM Implementation (Bridging UML and OWL),” <http://www.eclipse.org/m2m/atl/usecases/ODMImplementation/>, 2007.
- [10] D. L. McGuinness, F. V. Harmelen, “OWL Web Ontology Language-Overview,” <http://www.w3.org/TR/2004/REC-owl-features-20040210/>. W3C Recommendation 10 February 2004.
- [11] M. K. Smith, C. Welty, D. L. McGuinness, “OWL Web Ontology Language-Guide”, <http://www.w3.org/TR/2004/REC-owl-guide-20040210/>. W3C Recommendation 10 February 2004.
- [12] M. Dean, G. Schreiber, S. Bechhofer, F.V. Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, L. A. Stein, “OWL Web Ontology Language-Reference,” <http://www.w3.org/TR/2004/REC-owl-ref-20040210/>. W3C Recommendation 10 February 2004.
- [13] W3C OWL Working Group, “OWL 2 Web Ontology Language Document Overview,” <http://www.w3.org/TR/2009/REC-owl2-overview-20091027/>. W3C Recommendation 27 October 2009.
- [14] K. Baclawski, M. K. Kokar, P. A. Kogut, L. Hart, J. Smith, W. S. Holmes, J. Letkowski, M. L. Aronson, “Extending UML to Support Ontology Engineering for the Semantic Web,” (pp. 342-360). Springer Berlin Heidelberg, 2001.
- [15] D. Gašević, D. Djurić, V. Devedžić, V. Damjanović, “Converting UML to OWL Ontologies,” In : *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*. ACM, p. 488-489. 2004.
- [16] K. Kiko, C. Atkinson, “A Detailed Comparison of UML and OWL,” Technical Report, Reihe Informatik, TR-2008-004, 2008.
- [17] R. Bardohl, H. Ehrig, J. De Lara, G. Taentzer, “Integrating Meta Modelling with Graph Transformation for Efficient Visual Language Definition and Model Manipulation,” *Lecture Notes in Computer Science* 2984, pp.: 214-228. 2004.
- [18] H. Mouratidis, J. Odell, G. Manson, “Extending the Unified Modeling Language to Model Mobile Agents,” *Proceedings Agent Oriented Methodologies Workshop, Annual ACM Conference on Object Oriented Programming, Systems, Languages (OOPSLA)*, Seattle – USA, 2002.
- [19] K. Saleh, C. El-Morr, “M-UML: an extension to UML for the modeling of mobile agent-based software systems,” *Journal of Information and Software Technology*, ELSEVIER, Vol 46, 2004, pp. 219–227.
- [20] A. Belghiat, M. Bourahla, “An Approach based AToM3 for the Generation of OWL Ontologies from UML Diagrams,” *International Journal of Computer Applications* (0975 – 8887) Volume 41– No.3, March 2012.
- [21] E. Kerkouche and A. Chaoui, “A Formal Framework and a Tool for the Specification and Analysis of G-Nets Models Based on Graph Transformation,” *International Conference on Distributed Computing and Networking -CDCN’09-*, LNCS 5408, pp. 206–211, Springer-Verlag Berlin Heidelberg, India, 3-6 January, 2009.
- [22] F. Guerrouf, A. Chaoui, A. Aldahoud, “A graph transformation approach of mobile activity diagram to nested Petri nets,” *IJCAET* 5(1): 44-57 (2013).
- [23] M. R. Bahri, A. Hettab, A. Chaoui, and E. Kerkouche, “Transforming Mobile UML Statecharts Models to Nested Nets Models using Graph Grammars: An Approach for Modeling and Analysis of Mobile Agent-Based Software Systems,” in *SEEFM ‘09*, IEEE Computer Society Washington, pp.33-39, 2009.
- [24] SWOOP. Home page: <http://www.mindswap.org/2004/SWOOP/>.
- [25] E. Kerkouche, A. Chaoui, E. Bourennane, and O. Labbani, “Modeling and verification of dynamic behaviour in UML models, a graph transformation based approach,” *proceedings of SEDE’2009*, Las Vegas, Nevada, USA, 22-24 June 2009.
- [26] A. Belghiat, A. Chaoui, M. Maouche, M. Beldjehem, “Formalization of Mobile UML Statechart Diagrams using the π -calculus: An Approach for Modeling and Analysis,” In G. Dregvaite and R. Damasevicius (Eds.): *ICIST 2014*, CCIS 465, pp. 236–247. Springer.
- A. Schurr, “Specification of Graph Translators with Triple Graph Grammars,” In G. Tinhofer, editor, *WG’94 20th Int. Workshop on Graph-Theoretic Concepts in Computer Science*, volume 903 of *Lecture*


```

View Source - http://MobileUML.to.OWL#Mobile_CD_Ontology
File Edit Update Model
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE rdf:RDF [
  <ENTITY Mobile_CD_Ontology "http://MobileUML.to.OWL#Mobile_CD_Ontology">
  <ENTITY owl "http://www.w3.org/2002/07/owl#">
  <ENTITY rdfs "http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  <ENTITY xsd "http://www.w3.org/2001/XMLSchema#">
  <ENTITY xsd "http://www.w3.org/2001/XMLSchema#">
]
<rdf:RDF xmlns="http://Mobile_CD_Ontology;"
  xmlns:owl="owl:"
  xmlns:rdfs="rdfs:"
  xmlns:xsd="xsd:">
  <!-- Ontology Information -->
  <owl:Ontology rdf:about="#Mobile_CD_Ontology"
    rdfs:label="Written by Aissam BELGHIAI"/>
  <!-- Classes -->
  <owl:Class rdf:about="#Affected-VN">
    <rdfs:subClassOf>
      <owl:Restriction>
        <owl:minCardinality rdf:datatype="xsd:nonNegativeInteger">0</owl:minCardinality>
        <owl:onProperty rdf:resource="#Affected-VN.vcs"/>
      </owl:Restriction>
      <rdfs:subClassOf>
        <owl:Restriction>
          <owl:maxCardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:maxCardinality>
          <owl:onProperty rdf:resource="#Affected-VN.vcs"/>
        </owl:Restriction>
      </rdfs:subClassOf>
    </owl:Class>
  <owl:Class rdf:about="#Affected-V0">
    <rdfs:subClassOf>
      <owl:Restriction>
        <owl:cardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:cardinality>
        <owl:onProperty rdf:resource="#Affected-V0.vcs"/>
      </owl:Restriction>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:about="#Mobile-VC">
    <rdfs:subClassOf>
      <owl:Restriction>
        <owl:maxCardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:maxCardinality>
        <owl:onProperty rdf:resource="#Mobile-VC.vcs"/>
      </owl:Restriction>
      <rdfs:subClassOf>
        <owl:Restriction>
          <owl:cardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:cardinality>
          <owl:onProperty rdf:resource="#Mobile-VC.vms"/>
        </owl:Restriction>
      </rdfs:subClassOf>
      <rdfs:subClassOf>
        <owl:Restriction>
          <owl:cardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:cardinality>
          <owl:onProperty rdf:resource="#Mobile-VC.loggers"/>
        </owl:Restriction>
      </rdfs:subClassOf>
      <rdfs:subClassOf>
        <owl:Restriction>
          <owl:minCardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:minCardinality>
          <owl:onProperty rdf:resource="#Mobile-VC.vcs"/>
        </owl:Restriction>
      </rdfs:subClassOf>
    </owl:Class>
  <owl:Class rdf:about="#Remote-Logger">
    <rdfs:subClassOf>
      <owl:Restriction>
        <owl:cardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:cardinality>
        <owl:onProperty rdf:resource="#Remote-Logger.vcs"/>
      </owl:Restriction>
    </rdfs:subClassOf>
  </owl:Class>
  <!-- Datatypes -->
  <rdfs:Datatype rdf:about="xsd:string"/>
  <!-- Annotation Properties -->
  <owl:AnnotationProperty rdf:about="rdfs:label"/>
  <!-- Datatype Properties -->
  <owl:DatatypeProperty rdf:about="#Affected-VN.GetVotersList">
    <rdfs:domain rdf:resource="#Affected-VN"/>
    <rdfs:range rdf:resource="xsd:string"/>
  </owl:DatatypeProperty>
  <owl:DatatypeProperty rdf:about="#Affected-V0.GetVote">
    <rdfs:domain rdf:resource="#Affected-V0"/>
    <rdfs:range rdf:resource="xsd:string"/>
  </owl:DatatypeProperty>
  <owl:DatatypeProperty rdf:about="#Mobile-VC.RequestLogList">
    <rdfs:domain rdf:resource="#Mobile-VC"/>
    <rdfs:range rdf:resource="xsd:string"/>
  </owl:DatatypeProperty>
  <owl:DatatypeProperty rdf:about="#Remote-Logger.LogList">
    <rdfs:domain rdf:resource="#Remote-Logger"/>
    <rdfs:range rdf:resource="xsd:string"/>
  </owl:DatatypeProperty>
  <!-- Object Properties -->
  <owl:ObjectProperty rdf:about="#Affected-VN.vcs">
    <rdfs:domain rdf:resource="#Affected-VN"/>
    <rdfs:range rdf:resource="#Mobile-VC"/>
    <rdfs:subPropertyOf rdf:resource="#Inverse-contact"/>
  </owl:ObjectProperty>
  <owl:ObjectProperty rdf:about="#Affected-V0.vcs">
    <rdfs:domain rdf:resource="#Affected-V0"/>
    <rdfs:range rdf:resource="#Mobile-VC"/>
    <rdfs:subPropertyOf rdf:resource="#Inverse-visit"/>
  </owl:ObjectProperty>
  <owl:ObjectProperty rdf:about="#Inverse-contact">
    <rdfs:domain rdf:resource="#Affected-VN"/>
    <rdfs:range rdf:resource="#Mobile-VC"/>
    <owl:inverseOf rdf:resource="#contact"/>
  </owl:ObjectProperty>
  <owl:ObjectProperty rdf:about="#Inverse-log">
    <rdfs:domain rdf:resource="#Remote-Logger"/>
    <rdfs:range rdf:resource="#Mobile-VC"/>
    <owl:inverseOf rdf:resource="#log"/>
  </owl:ObjectProperty>
  <owl:ObjectProperty rdf:about="#Inverse-visit">
    <rdfs:domain rdf:resource="#Affected-V0"/>
    <rdfs:range rdf:resource="#Mobile-VC"/>
    <owl:inverseOf rdf:resource="#visit"/>
  </owl:ObjectProperty>
  <owl:ObjectProperty rdf:about="#Mobile-VC.loggers">
    <rdfs:domain rdf:resource="#Mobile-VC"/>
    <rdfs:range rdf:resource="#Remote-Logger"/>
    <rdfs:subPropertyOf rdf:resource="#log"/>
  </owl:ObjectProperty>
  <owl:ObjectProperty rdf:about="#Mobile-VC.vms">
    <rdfs:domain rdf:resource="#Mobile-VC"/>
    <rdfs:range rdf:resource="#Affected-VN"/>
    <rdfs:subPropertyOf rdf:resource="#contact"/>
  </owl:ObjectProperty>
  <owl:ObjectProperty rdf:about="#Mobile-VC.vcs">
    <rdfs:domain rdf:resource="#Mobile-VC"/>
    <rdfs:range rdf:resource="#Affected-V0"/>
    <rdfs:subPropertyOf rdf:resource="#visit"/>
  </owl:ObjectProperty>
  <owl:ObjectProperty rdf:about="#Remote-Logger.vcs">
    <rdfs:domain rdf:resource="#Remote-Logger"/>
    <rdfs:range rdf:resource="#Mobile-VC"/>
    <rdfs:subPropertyOf rdf:resource="#Inverse-log"/>
  </owl:ObjectProperty>
  <owl:ObjectProperty rdf:about="#contact">
    <rdfs:domain rdf:resource="#Mobile-VC"/>
    <rdfs:range rdf:resource="#Affected-VN"/>
    <owl:inverseOf rdf:resource="#Inverse-contact"/>
  </owl:ObjectProperty>
  <owl:ObjectProperty rdf:about="#log">
    <rdfs:domain rdf:resource="#Mobile-VC"/>
    <rdfs:range rdf:resource="#Remote-Logger"/>
    <owl:inverseOf rdf:resource="#Inverse-log"/>
  </owl:ObjectProperty>
  <owl:ObjectProperty rdf:about="#visit">
    <rdfs:domain rdf:resource="#Mobile-VC"/>
    <rdfs:range rdf:resource="#Affected-V0"/>
    <owl:inverseOf rdf:resource="#Inverse-visit"/>
  </owl:ObjectProperty>
  </rdf:RDF>
  
```

Fig. 11. The generated OWL ontology.

Artificial Bee Colony Based Focus Fusion

Veysel Aslantas, Ahmet Nusret Toprak
Erciyes University, Computer Engineering Department
Kayseri, TURKEY
aslantas@erciyes.edu.tr, antoprak@erciyes.edu.tr

Abstract— This paper proposes a pixel based multi-focus image fusion method which consists of two main steps: detection of point spread functions of the source images by using artificial bee colony algorithm and fusion of the source image by making use of estimated point spread functions. The main advantage of the proposed method is it directly fuses the source images without any transformation that modifies the original pixel values. Experimental results demonstrate the superiority of the proposed method in terms of quantitative and visual evolution.

Keywords—multi-focus image fusion; artificial bee colony, blur estimation

I. INTRODUCTION

Images taken by digital cameras usually suffer from limited depth of field (DoF) problem. Limited DoF disallows the digital cameras to take all-in-focus images. Therefore the obtained image appears partially blurred. Multi-focus image fusion techniques solve this problem by taking multiple images with different focus points and fusing them together to obtain an all-in-focus image [1]. Multi-focus image fusion techniques have been used widely in digital camera and microscopy applications.

The multi-focus image fusion methods generally divided into two groups: spatial domain methods are based on direct fusion of pixels of the input images; transform domain processing methods are based on fusing the transforms of the images [2].

Transform domain based methods consist of three main steps: applying a transform to the each input images, fusing all transforms to produce the fused representation, and at last applying the inverse transform to the fused representation to obtain the final fused image. The most generally used transform based methods can be listed as Laplacian Pyramid (LP) [3], Discrete Wavelet Transform (DWT) [4], and Discrete Cosine Transform (DCT) [5]. However, applying a transform prior to fusion modifies the original pixels of input images. Therefore, brightness and color distortions may occur on the fused image [6].

Different from the transform domain based methods, spatial domain based methods directly fuse the input images by taking their spatial features into account. These methods search for sharp pixels or regions to be transferred in order to construct the fused image. Spatial domain based divided into two main groups: region based and pixel based methods. The former methods such as block selection method [7], Differential Evolution based block selection [8], region selection [9], initially segment the input images by a segmentation algorithm and then produce fused image by selecting the sharper regions.

The latter methods such as spatial frequency based method [10] determine the sharp pixels that form the fused image.

In this paper, a pixel based spatial domain method is proposed. The proposed method has two main steps: estimating the point spread functions (PSF) of input image and fusing input images by using the estimated PSFs. In the multi-focus image fusion applications, a collection of images with different focus points is captured. Each image of this collection includes both blurred and sharp regions. While a particular object in the scene appears in focus in one image, it appears out of focus in others. In other words, both blurred and sharp images of each object in the scene exist in the collection. Blurred image of an object can be defined as the convolution of the same object with a particular PSF [11]. This means if the both blurred and sharp images exist, then the PSF can be calculated. However, there is no prior knowledge about which region of the input images are blurred or sharp. Therefore a novel technique that using ABC to estimate the PSFs of the input images is proposed in this paper. ABC is simple yet effective. Therefore it is straightforward to solve optimization problems. ABC has been also employed to solve several problems in image processing area. A good review can be found in [12]. The proposed method, firstly calculate the PSFs of the input images by using ABC. Then, it detects sharp pixels by making use of estimated PSFs. Finally, it forms the fused image by transferring the detected sharp pixels.

The rest of paper is organized as follows. Section II explains the proposed method in detail. Experimental results are given in Section III. Finally, Section IV concludes the paper.

II. FOCUS FUSION BASED ON ARTIFICIAL BEE COLONY

In this section, firstly, the ABC algorithm is briefly described. Then, the ABC based PSF estimation and focus fusion methods are introduced.

A. Artificial Bee Colony (ABC) Algorithm

Artificial Bee Colony (ABC) is a swarm based meta-heuristic global optimization algorithm that was defined by Karaboga [13]. It is motivated by the intelligent foraging behavior of the honey bees. ABC has only three control parameters: population size, maximum cycle number and limit.

In ABC algorithm, each possible solution is represented by a food source and the nectar amount of the food sources represent the quality of the corresponding solution. The artificial bee colony consists of three groups of bees: employed bees which are employed at a specific food source, onlooker bees that watch the dance of employed bees to choose a food source, and scout bees that search the environment randomly. Both of the onlookers and the scouts are also called as unemployed foragers. Initially, all food sources are explored by scout bees. Then, the nectar of food sources are exploited by employed and onlooker bees and this cause them to become exhausted. Afterwards, the employed bees whose food source is exhausted become a scout bee. The number of the employed bees is equal to the number of food sources (solutions).

At the first step, the initial population of the SN food sources is initialized by the scout bees where SN denotes the number of employed bees. Each food source $x_m(m=1,2,\dots,SN)$ is a solution vector with n variables $x_{mi}(i=1,2,\dots,n)$ which are the optimization parameters. In order to initialize the food sources the following formula is used.

$$x_{mi} = l_i + rand(0,1) * (u_i - l_i) \quad (1)$$

where l_i and u_i are the lower and the upper bounds, respectively.

After the initialization, there are three phases of ABC algorithm. First one is employed bees phase. Employed bees search for the new food sources (v_m) that have more nectar, in the neighborhood of the food source (x_m) in their memory. They determine a neighbor food source using the following definition.

$$v_{mi} = x_{mi} + \phi_{mi}(x_{mi} - x_{ki}) \quad (2)$$

where x_k is randomly chosen food source, i is a randomly selected parameter, ϕ is a random number. After producing the new food source v_m , its fitness value is calculated and the food source that has the best fitness value is selected between x_m and v_m .

Second phase is the onlooker bees phase. As mentioned above, there are two kind of unemployed bees: onlooker and scout bees. Employed bees share the information about their food source with the onlooker bees. Then onlooker bees choose their food source by considering the probability value that calculated using the fitness value providing by employed bees. The probability value p_m is calculated by using the following expression.

$$p_m = \frac{f(x_m)}{\sum_{m=1}^{SN} f(x_m)} \quad (3)$$

where f is the fitness function.

After food source (x_m) is selected, a neighbor food source v_m is found out by using the eq. (2). As in the previous phase, a greedy selection is applied between x_m and v_m .

The last phase is the scout bees phase. Scout bees are the unemployed bees that search their food sources randomly. If an employed bee cannot improve its solution through a predefined number of iterations called *limit*, it becomes scout bee and its solutions are deserted. Thereafter, converted scout bee start to search for new food source by using (1).

B. ABC Based PSF Estimation

In this section a novel method that uses ABC to estimate the PSF function of the multi-focus input image is presented. Each input image of multi-focus collection includes both focused (sharp) and defocused (blurred) regions. For the simplicity, assume that collection contains two input images. Input images (I_1 and I_2) can be defined by following definitions.

$$\begin{aligned} I_1 &= f_1(x, y) + g_1(x, y) \\ I_2 &= f_2(x, y) + g_2(x, y) \end{aligned} \quad (4)$$

where $f(x, y)$ and $g(x, y)$ represent the focused and defocused pixels of input images, respectively. Since defocused image of an object is equal to convolution of the focused image of same object with a particular PSF, defocused image $g(x, y)$ can be written as following equation.

$$g(x, y) = f(x, y) \otimes h(x, y, \sigma) \quad (5)$$

where $h(x, y, \sigma)$ is the PSF function. The PSF of an imaging system describes the shape of the blur formed when a point object is imaged. PSF best fit to a 2D Gaussian model [14]:

$$h(x, y) = (1 / 2\pi\sigma^2) \exp(-(x^2 + y^2) / 2\sigma^2) \quad (6)$$

where σ denotes the spread parameter (SP).

Proposed PSF estimation method employs the ABC in order to estimate the PSFs of the input images. As can be seen from (6), the only parameter of the PSF is spread parameter (σ). Thus, SPs of the input images are searched for each input image in the optimization process. In our example, each individuals of ABC consist of two parameters that are the SPs (σ_1 and σ_2) of PSFs of the input images (h_1 and h_2).

Initially, the control parameters of the ABC maximum cycle number, population size and limit are defined. Then initial population is generated randomly by using (1).

In the each cycle of ABC, for each individual following steps are repeated until stopping criteria is met. First, fitness value of the each individual is calculated. To this end, PSFs are produced by substituting spread parameter of each individual in (6). Artificially blurred images are produced by convolving each input image with the other's PSF. In our example, I_1 is convolved with h_2 and I_2 is convolved with h_1 to generate artificially blurred images \hat{I}_1 and \hat{I}_2 , respectively.

$$\begin{aligned} I_1(x, y) &= I_1(x, y) \otimes h_2(x, y, \sigma_1) \\ I_2(x, y) &= I_2(x, y) \otimes h_1(x, y, \sigma_2) \end{aligned} \quad (7)$$

Then, difference images (D_1 and D_2) are obtained by subtracting I_1 from \hat{I}_2 and I_2 from \hat{I}_1 .

$$\begin{aligned} D_1 &= |I_1 - I_2| \\ D_2 &= |I_2 - I_1| \end{aligned} \quad (8)$$

If the optimum PSFs are obtained, the dot product of the difference image should be equal to zero. Thus the following definition is used as the fitness function.

In order to generate the next generation of the population, employed bees, onlooker bees and scout bees phases are executed, respectively. If the stopping criterion is not met, the above steps are repeated.

C. Fusion of Input Images Using Optimal PSFs

After the estimation of optimal PSFs of the input images, sharp images can be determined. First artificially blurred images (\hat{I}_1 and \hat{I}_2) are produced by substituting optimally estimated PSFs in (7). Since the PSFs that are used to generate the artificially blurred images are optimal, blurred regions of input images are obtained blurred with same degree in artificially blurred ones. In other words, pixel values of the blurred regions of the input images are equal to values of the corresponding pixels in the artificially blurred ones.

Once the artificially blurred images are produced, difference images (D_1 and D_2) can be calculated by using (8). Since optimal PSFs are used, blurred regions ($g(x, y)$) of input images (I_1 and I_2) are canceled out in difference images (D_1 and D_2), respectively. In this manner, each pixel of the difference images is compared as following expression.

$$I_F(x, y) = \begin{cases} I_1(x, y) & D_1(x, y) \geq D_2(x, y) \\ I_2(x, y) & D_1(x, y) < D_2(x, y) \end{cases} \quad (9)$$

Fused image (I_F) is generated by transferring pixels from input image which corresponding value of difference image is bigger.

III. EXPERIMENTAL RESULTS

In this section the visual and quantitative evaluation about fusion performance of the proposed method is reported. The results of the proposed method are compared with three well-known methods: Discrete Wavelet Transform (DWT) [4], Discrete Cosine Transform (DCT) [5] and Pixel based method (PB) [10]. The parameters of the methods are selected as: for DWT filter is "Sym8" and decomposition level is 7; for PB sharpness value is spatial frequency (SF). For the proposed method, as a result of many experiments control parameter of ABC is selected as: NP=16, L=20.

To assess the performance of the fusion methods quantitatively, three objective quality metric: quality of edges (Q_E) [15], mutual information (Q_{MI}) [16] and structural

similarity (Q_{SSIM}) [17] are used. Experiments are conducted on three different multi-focus image sets: *Clock*, *Matches* and *Watch*. The input images of these sets are illustrated in Fig. 1.

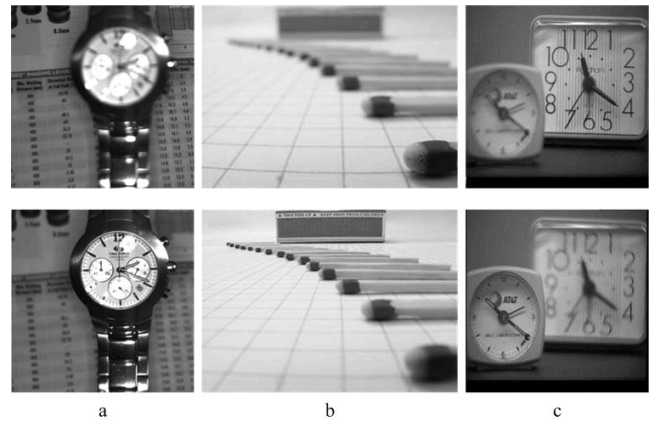


Fig. 1. Input image sets: (a) Watch, (b) Matches, (c) Clock.

The first experiment is carried out on the *Watch* image set. In this image set there is a watch on a paper. In the first image watch is focused and in the other background is focused. The fusion results and the difference images that are obtained by subtracting fused image from each input image, of the *Watch* image set is given in Fig. 2. The small parts of the watch make this image set challenging. As can be seen from the Fig. 2. transform domain based methods DWT and DCT produce artifacts around the detail parts of watch. This artifact can be easily seen in difference images of these methods. Although there are also some minor artifacts in the fused image of PB method, it introduces an acceptable result. By contrast, the proposed method does not produce any artifacts and well preserve detail parts of the watch.

The experiments are also performed on the *Matches* image set. In this image set match sticks are arranged in order and a matchbox is also located at the background of the images. Fused and difference images of the fusion methods are shown in the Fig. 3. As can be observed from the images DCT method produces reasonable result. However, first match is blurred in the fused image. On the other hand, some major artifacts exist in the results of DWT and PB methods. These especially cannot preserve the match heads. On the contrary, proposed method well preserves matches and the matchbox on the fuse image.

The last experiments are conducted on the *Clock* image set. In this image set there are two clocks on a table. In the each image one of the clocks is focused. The fusion results of the *Clock* image set is given in Fig. 4. Difference images of DWT reveal that DWT cannot well preserve the details of the clocks and modifies the original pixel values. On the other hand DCT and PB methods introduce blocking artifacts. By contrast, the proposed method can effectively fuse the images and preserves the details.

The objective performance of the methods is also evaluated. Table 1 gives the objective quality values of the methods in terms of Q_E , Q_{MI} and Q_{SSIM} metrics. From the table, it can be seen that the proposed method provides the best fusion results in terms of Q_E , Q_{MI} and Q_{SSIM} metrics. For Clock image DCT produce a better Q_{MI} result. However this results does not coincide with the visual result.

IV. CONCLUSION

In this paper, an effective multi-focus image fusion method is proposed. Furthermore, a method that estimates the PSFs of the multi-focus input images by using ABC is also introduced. The proposed multi-focus image fusion method initially estimates the PSFs of the input images. Then each input image is artificially blurred by convolving estimated PSFs. By using these artificially blurred images, sharp pixels of the input images are determined. Eventually, the all-in-focus fused image is generated by gathering the sharp pixels. Experiments show that the proposed method can produce fused images that have high visual quality. Moreover, the effectiveness of the proposed method is demonstrated by quantitative metrics.

TABLE I. QUANTITATIVE ASSESMENT OF FUSION METHODS

Images	Metrics	Methods			
		DWT	DCT	PB	Proposed
Watch	Q_E	0,6488	0,7167	0,7217	0,7226
	Q_{MI}	5,6141	9,3345	9,2185	9,3398
	Q_{SSIM}	0,9514	0,9766	0,9755	0,9771
Matches	Q_E	0,7215	0,7658	0,7826	0,7895
	Q_{MI}	7,6860	12,5990	12,3110	12,3948
	Q_{SSIM}	0,9810	0,9753	0,9834	0,9852
Clock	Q_E	0,6710	0,7387	0,7386	0,7388
	Q_{MI}	5,8839	9,0800	9,0406	9,0413
	Q_{SSIM}	0,9590	0,9806	0,9790	0,9812

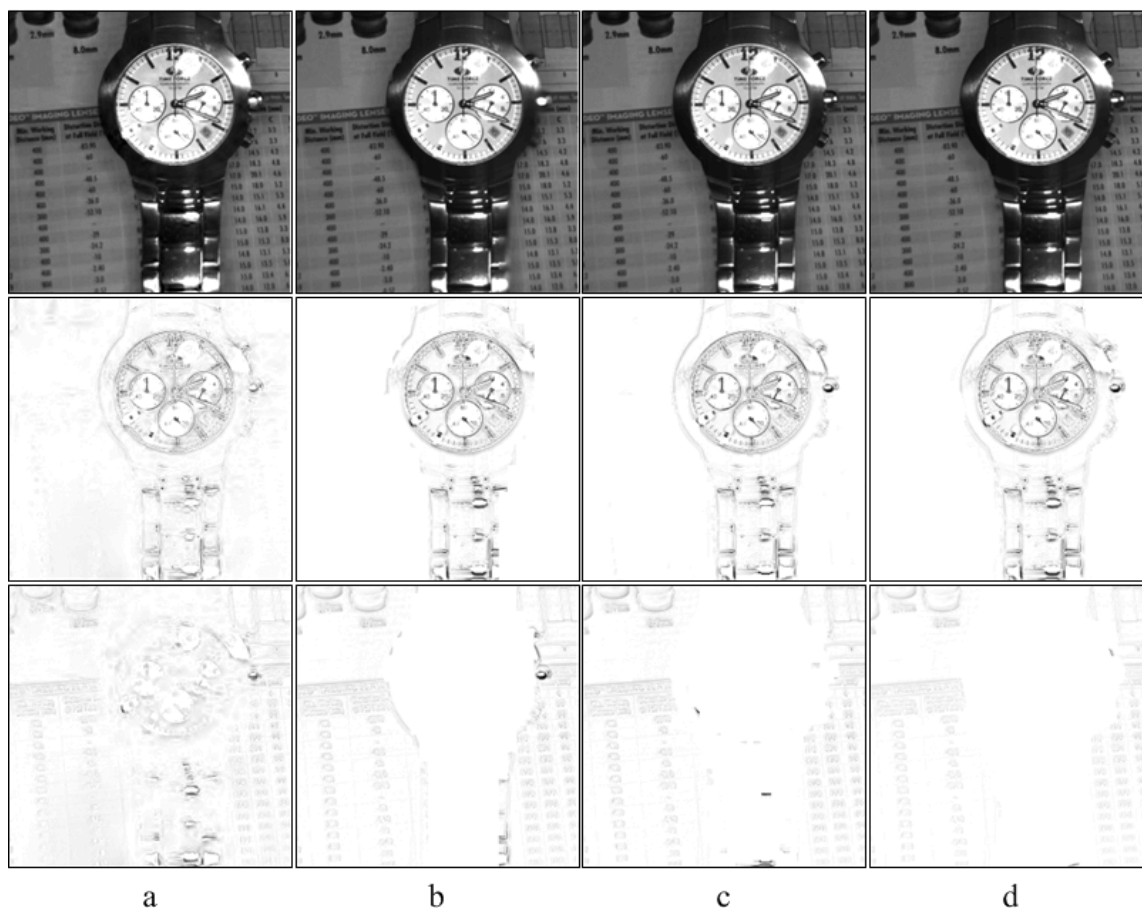


Fig. 2. Fused and difference images of the Watch image set: (a) DWT (b) DCT (c) PB (d) Proposed method.

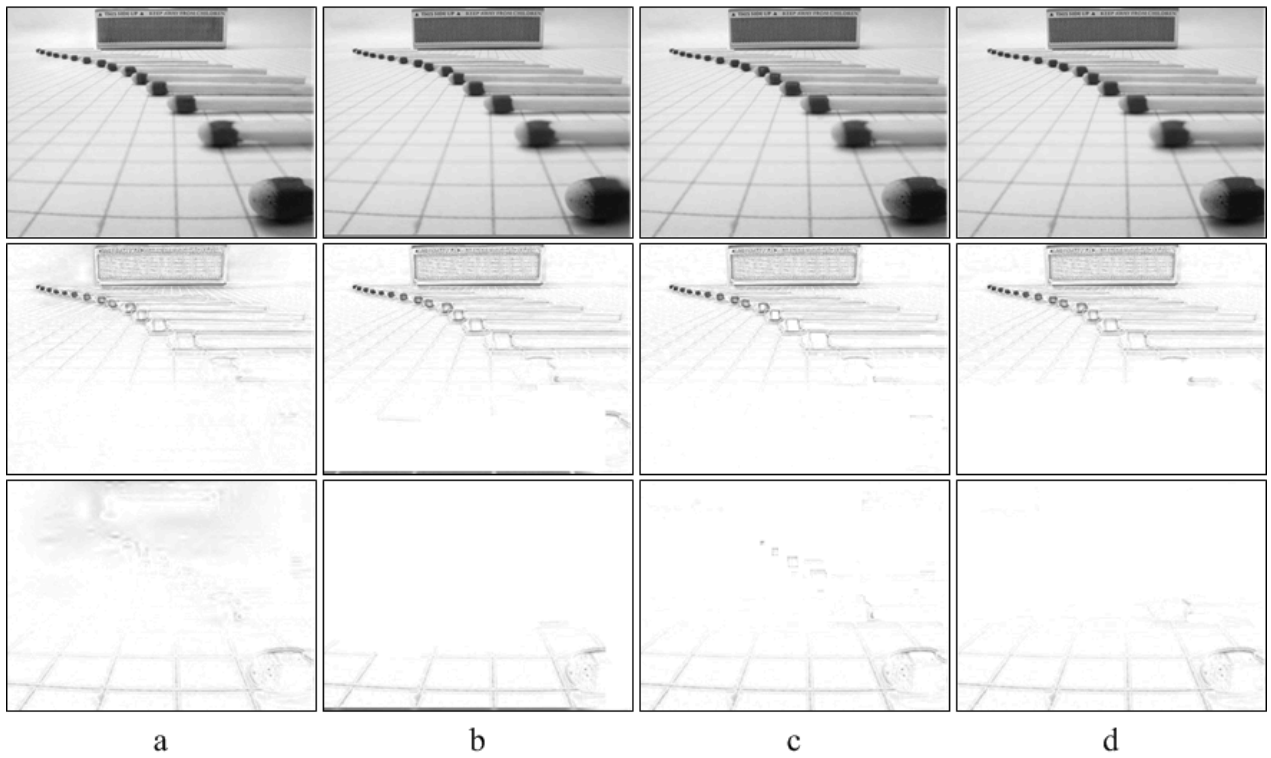


Fig. 3. Fused and difference images of the Matches image set: (a) DWT (b) DCT (c) PB (d) Proposed method.

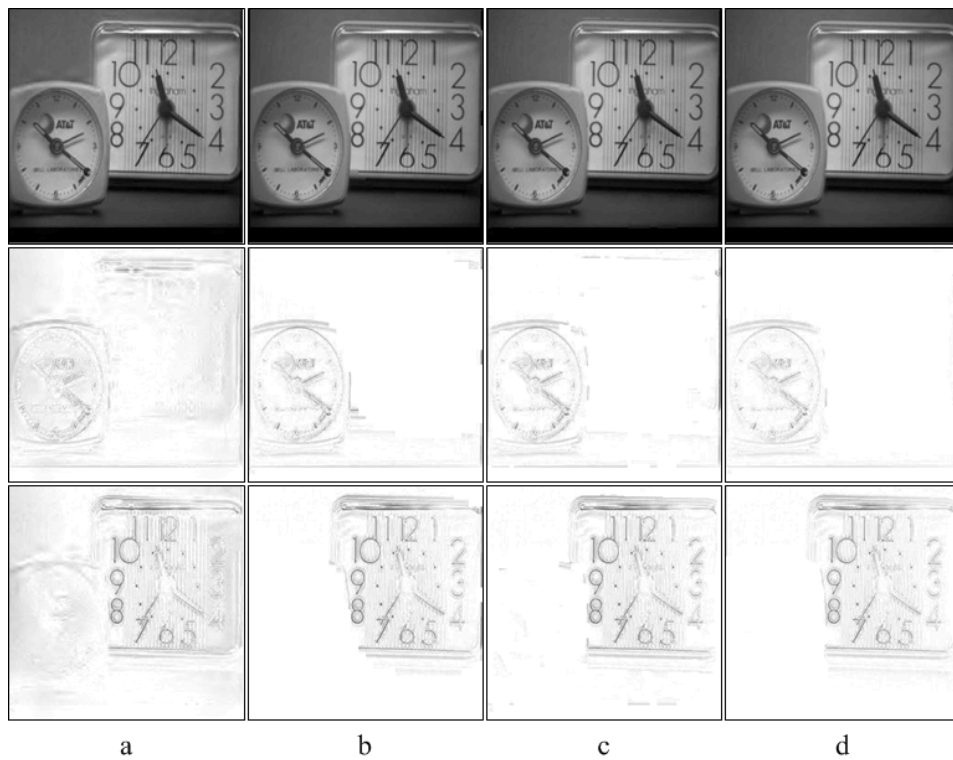


Fig. 4. Fused and difference images of the Clock image set: (a) DWT (b) DCT (c) PB (d) Proposed method

ACKNOWLEDGMENT

This paper is supported by Research Foundation of the Erciyes University, Kayseri, Turkey (Grant no: FDK-2015-5630).

REFERENCES

- [1] V. Aslantas and A. N. Toprak, "A pixel based multi-focus image fusion method," *Optics Communications*, vol. 332, pp. 350-358, 2014.
- [2] V. Aslantas, E. Bendes, R. Kurban, and A. N. Toprak, "New optimised region-based multi-scale image fusion method for thermal and visible images," *Image Processing, IET*, vol. 8, pp. 289-299, 2014.
- [3] P. J. Burt and R. J. Kolczynski, "Enhanced image capture through fusion," in *Computer Vision, 1993. Proceedings., Fourth International Conference on*, 1993, pp. 173-182.
- [4] G. Pajares and J. M. de la Cruz, "A wavelet-based image fusion tutorial," *Pattern Recognition*, vol. 37, pp. 1855-1872, Sep 2004.
- [5] M. B. A. Haghghat, A. Aghagolzadeh, and H. Seyedarabi, "Multi-focus image fusion for visual sensor networks in DCT domain," *Computers & Electrical Engineering*, vol. 37, pp. 789-797, 2011.
- [6] S. Li, X. Kang, and J. Hu, "Image Fusion With Guided Filtering," *Image Processing, IEEE Transactions on*, vol. 22, pp. 2864-2875, 2013.
- [7] S. Li, J. T. Kwok, and Y. Wang, "Combination of images with diverse focuses using the spatial frequency," *Information Fusion*, vol. 2, pp. 169-176, 2001.
- [8] V. Aslantas and R. Kurban, "Fusion of multi-focus images using differential evolution algorithm," *Expert System Applications*, vol. 37, pp. 8861-8870, Dec 2010.
- [9] S. Li and B. Yang, "Multifocus image fusion using region segmentation and spatial frequency," *Image and Vision Computing*, vol. 26, pp. 971-979, 2008.
- [10] B. Yang and S. Li, "Multi-focus image fusion based on spatial frequency and morphological operators," *Chinese Optics Letters*, vol. 5, pp. 452-453, 2007.
- [11] V. Aslantas, "A depth estimation algorithm with a single image," *Optics Express*, vol. 15, pp. 5024-5029, 2007.
- [12] D. Karaboga, B. Gorkemli, C. Ozturk, and N. Karaboga, "A comprehensive survey: Artificial bee colony (ABC) algorithm and applications," *Artificial Intelligence Review*, vol. 42, pp. 21-57, 2014.
- [13] D. Karaboga and B. Basturk, "A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm," *Journal of Global Optimization*, vol. 39, pp. 459-471, 2007/11/01 2007.
- [14] M. Subbarao, T. C. Wei, and G. Surya, "Focused image recovery from two defocused images recorded with different camera settings," *Image Processing, IEEE Transactions on*, vol. 4, pp. 1613-1628, 1995.
- [15] C. S. Xydeas and V. Petrovic, "Objective image fusion performance measure," *Electronics Letters*, vol. 36, pp. 308-309, Feb 2000.
- [16] Q. Guihong, Z. Dali, and Y. Pingfan, "Information measure for performance of image fusion," *Electronics Letters*, vol. 38, pp. 313-315, 2002.
- [17] C. Yang, J.-Q. Zhang, X.-R. Wang, and X. Liu, "A novel similarity based quality metric for image fusion," *Information Fusion*, vol. 9, pp. 156-160, 2008.

Ontology-based knowledge recognition in service-oriented virtual research environments

Case: application in e-learning

Anatoly Gladun

*International Research and Training Center of Information Technologies and Systems of National Academy of Sciences Ukraine,
Kyiv, Ukraine
glanat@yahoo.com*

Julia Rogushina

*Institute of Software Systems of National Academy of Sciences Ukraine
Kyiv, Ukraine
ladamandraka2010@gmail.com*

Jeanne Schreurs

*Department of Business Informatics, Professor Emeritus in Hasselt University
Hasselt, Belgium
jeanne.schreurs@uhasselt.be*

Abdel-Badeeh Salem

*Department of Computer Science, Head of Artificial Intelligence and Knowledge Engineering Research Labs, Faculty of
Computer and Information Sciences, Ain Shams University
Cairo, Egypt
abmsalem@yahoo.com*

Abstract—In the paper some methods of ontology-based knowledge recognition in service-oriented virtual research environment are proposed. These methods are about export of knowledge, qualification level and study domain of students, and about automatic evaluation of their skills. The research is situated in different disciplines. Using domain ontology as an instrument for student skills evaluation is set forward. Web services and ontologies provide reuse of these methods in other applications. A prototype automatic tutor has been developed to support e-learning.

Keywords—*knowledge management, intelligent e-learning, ontology, virtual research environment, knowledge recognition*

I. INTRODUCTION

Virtual Research Environments are providing a lot of possibilities for distributed knowledge management to be applied in different study domains. Implementing knowledge management in modern universities is a challenge when they are providing a mix of traditional and on distance education. Knowledge management process can be organised in different ways. The following steps are often identified: acquisition, creation, storage, validation, and utilisation of knowledge. These steps can be found in e-learning projects setup to increase the learning process effectiveness.

Knowledge acquisition about the qualification level and the learners' skills is a main problem. This problem can be seen as a particular case of pattern recognition. The information object describes the qualification and the skills of the learners. An approach based on ontologies is widely used for solving these problems. In our research we are proposing a method of reference domain ontology to be used as an instrument to

evaluate students' qualification and skills. The students- or course-ontology is compared with the reference one based on a set of different concepts and relation ratings.

II. VIRTUAL RESEARCH ENVIRONMENTS

We believe that more research about generating new knowledge and cost-effective technologies, mainly based on a number of ICT-related disciplines, will offer a number of possibilities, which have not been exploited yet in Virtual Research Environments (VREs) supported by e-infrastructures. In particular, state-of-the-art methods and technologies in fields like the Semantic Web, Computing, Networks, Artificial Intelligence, among others, will be integrated into the SMART-VRE solution.

We are analyzing in this research only the VRE functionality on top of real use cases, and by the way make it possible to take into account the privacy aspects. The communication and dissemination strategy of the VRE have a

key role into the accomplishment of its main technical objectives by: (1) reporting to Universities and Research Institutes, to the general public and to the media; (2) exploiting the VRE outcomes and results in order to help reinforcing the EU industrial base in the domain of e-infrastructures; (3) communicating about SMART-VRE benefits in VRE's so as to ensure the exploitation.

In table 1 we present an overview of the challenges faced by the implementation of a virtual research environment.

TABLE I. CHALLENGE FACING

Challenge	Way(s) to be faced
Integrate resources across all layers of the e-infrastructure (networking, computing, data, software, user interfaces)	<ul style="list-style-type: none"> -Adopting open data (fulfilling privacy requirements), open science and open innovation as main principles and implementing an advanced dedicated software application to facilitate e-infrastructure networking resources integration. -Encompassing several physical e-infrastructures and computing models, including HPC, grid and cloud computing models. -Performing semantic annotation of data for further semantic integration into ontologies using standardized ontological languages. - Using semantic web services and intelligent agents for integrating software applications. -Adopting a bottom-up approach in user interfaces integration, so achieving High-Fidelity prototypes of user interfaces that will reflect the scalability features used in the previous stage (Low-Fidelity prototypes)
Foster cross-disciplinary data interoperability	<ul style="list-style-type: none"> -Data will be semantically annotated so that these can be interoperated amongst VRE (web) services and users overcoming possible disciplinary-related terminological discrepancies. - Semantic web services will be utilized so that VRE-provided services and resources are decoupled with respect to both the data provided by such services and the (user) services requested
Provide functions allowing data citation and promoting data sharing and trust	Metadata will be semantically annotated for each data for those ones further processing so that they will include features like authorship and source of publication
Provide functions promoting data sharing	Maximizing the use of ontologies and semantic web services in carrying out the (services, networking and joint research) activities in the VRE platform
VREs should provide functions promoting trust	<ul style="list-style-type: none"> - We will seek endorsement of the SMART-VRE privacy concepts by consumer stakeholders and propose an European privacy standard for VRE solutions. - The developed software modules integrating the VRE platform in the project will be continually tested and user-evaluated. - The VRE platform will encompass security mechanisms and protocols against external attacks.

The overall aim of the VRE is the generation, validation, communication and exploitation of the VRE platform for ageing. In particular, the VRE platform will be conceived in such a way that: (1) it will be conceptually defined on a set of underlying ageing-relevant e-infrastructures; (2) it will re-use existing project theme-relevant knowledge and solutions (e.g.,

tools and services from existing infrastructures and projects) at both European and national levels; (3) standardized software building blocks and workflows, well-documented APIs and interoperable software components will be used for designing and implementing the VRE; (4) at least 1.000 potential users will be targeted.

The VRE platform manages data in such a way that their corresponding metadata semantics will be formally defined in a machine-understandable and interoperable manner. They will support proof of concept, prototyping and deployment of advanced data services and environments, and access to top-of-the-range connectivity and computing.

III. RESEARCH AND INNOVATION ACTIVITIES OF THE VRE

The following main types of research- and innovation-activities, covering a variety of research topics about the trans-disciplinary nature of the VRE, have been linked to the problem and the resulting solution.

A. Computer networks

1. High Performance Computation

High Performance Computation (HPC) is set forward. The current e-infrastructure services related to HPC, Grid and Cloud, which have been funded by national or European funding agencies (like FP7 PRACE for HPC, EGI-Inspire for Grid, BonFIRE for Cloud services), are focused on computational intensive services, rather than on data processing [1], [7].

As underlined in the PRACE report (“The scientific case for high-performance computing in Europe 2012-2020”), handling large data volumes generated by research is a major challenge and opportunity for future HPC systems and integrated environments for computing and data management. SMART-VRE intends to provide a showcase of an integrated environment that can serve a specific community, the one engaged in ageing research. Offering HPC services to various research communities is and was subject of multiple e-infrastructure projects funded by EC. The most remarkable ones are the communities around the PRACE initiative. The UVT team has offered HPC services in multiple EC projects (starting with the early FP6-Infra SCIENCE, for symbolic computing community until the latest FP7-eInfra HP-SEE, for computation physics, computational chemistry and life sciences). SMART-VRE is offering the opportunity to show how a particular health community can benefit from the availability of HPC resources.

Since specialized data services are becoming complex and expensive to maintain by datacenter management, a recent trend is their deployment in Private or Public Clouds. The migration and deployment is nowadays not straightforward and requires specific knowledge and manual intervention. [8],[9]

2. Networking.

Networking, or co-sharing computing services [10], is fostering forms of shared information thanks to the engagement of agents and resources improving participatory approaches and direct involvement. Networking is also critical to enforce and materialize the interrelations between innovation and processes of change whose role have been widely acknowledged and studied in literature. Dynamics and impacts of collaborative systems may also highly vary according to the action of varieties of well-known pathologies in social systems creating specific peculiarities of these networks. These pathologies have the potential capability of creating profound effects in inhibiting link formation, to turn positive links into ineffective or negative ones and to enhance the non-linear system behavior. And as a result these pathologies are deeply influencing the quality of the interactions among network agents. The possibility of providing a correct diagnosis of these network pathologies can alert about actual and potential possibility of the occurrence of a system collapse caused by deterioration in the link value and in the eventual link losses. It can also support in preventing a system collapse.

B. Data management

1. Open Science and Open Innovation.

Open Science and Open Innovation are key concepts, which have become very popular in the last years [11],[12].

Open Science refers to dynamic systems of knowledge production, characterized by a more or less high degree of accessibility of information and by knowledge of researchers and scientists. These systems act as dynamos, generators and stimulators of knowledge for future research. Open Science implies the creation of effective networks based on shared collaborative resources using technical tools that are able to distribute the information. The collaborative technologies are facilitating also the distribution resources including protected data (proprietary data and materials, trade secrets, legal protections, intellectual property rights, patents, copyright, etc.).

The Open Innovation concept is one of the central aspects of the processes of diffusion of innovation and technology transfer. This concept involves many disciplines including economics, psychology, sociology, cultural anthropology and management. In general, Open Innovation can be defined as the result of the use of purposive inflows and outflows of knowledge to accelerate internal innovation, and to expand the markets for external use of innovation. In literature, several international case studies are cited from which it is possible to understand the concrete operation of these processes and to identify the most important factors involved.

Both concepts of open innovation and of open science will guide the high-level strategy to carry out the networking activities in SMART-VRE (fig.1).

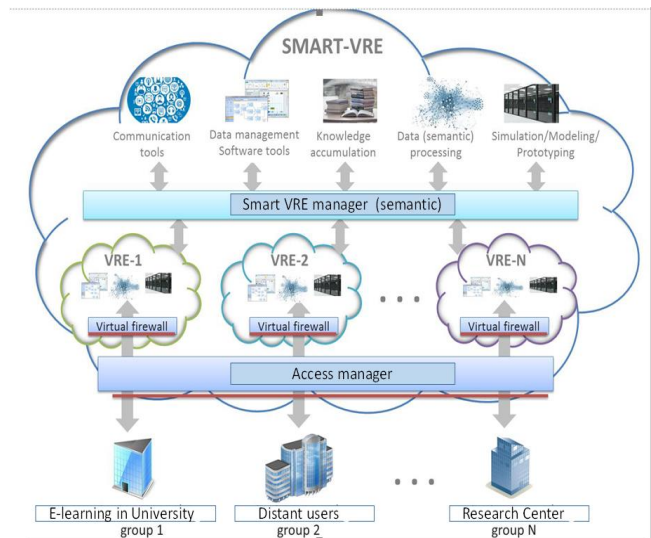


Fig. 1. The VRE Framework

2. Data management and Semantic Web technologies

To ensure the exploitation of data, data must be available and accessible in a network environment. However, the nature of data (research, administrative, academic) is variable and dependent of the scientific discipline, the application scope and the life cycle. A critical point in data management is the metadata representation of datasets catalogs [13] for which the vocabulary DCAT [14] is used. From a technical point of view, an open dataset has a life cycle that includes data extraction, storage, review, interconnection with other open data, classification and maintenance [15].

The correct management of research data is a fundamental part of the research process. This management involves making decisions and actions before the creation of the data, during its creation and use and throughout its life cycle. Management of data should involve 5 actions.

1. plan of data management, as part of the budgets of the organization, that anticipates management challenges and that proposes solutions to them;
2. treat of appropriate ethical and legal issues relating to sensitive personal data, to copyright and to license about access and use of data;
3. the organization and documentation of data according to disciplinary and international standards that allows to know the nature of the data and how the data was created and how it can be reused;
4. the appropriate storage, back-up and security mechanisms to ensure the confidentiality, integrity and availability of information;
5. standards about sharing the data when cited
6. archiving of a final copy of the data in specialized services, taking the necessary measures for its preservation and dissemination.

All these steps will be realized in a data management policy, which will be adopted in this proposal.

3. Ontologies.

Another analytic perspective of data management comes from their conceptual dimension. Conceptual systems, which are typically represented by concepts and categories, can be modeled by universal constraints independent of cultural variations [7], in which case the quality of the categorizations is positively correlated with the level of simplicity of these categorizations [9, 16].

Ontologies, which are commonly conceived as explicit formalizations of shared conceptual systems [17], are the most widely used approach to represent knowledge, due to their intrinsic properties of structure, reuse, sharing and formalization. All these properties enable them even for the automatic integration of knowledge once this has been represented [18]. Ontologies provide a common vocabulary of an area and define – with different levels of formality - the meaning of the terms and the relations between them. Knowledge in ontologies is mainly formalized using classes, relations, functions, axioms and instances [19].

C. Semantic Web

1. Ontologies to add semantics to the data on the web

Ontologies form the backbone on which to build the future Web, namely, the Semantic Web [20],[21].

Ontologies and reasoning techniques are leading to the achievement of a more intelligent Web [9] or to the automation of science [13]. The purpose of the Semantic Web (SW) is to add semantics to the data on the Web (for example, establish the meaning of the data using metadata), so that machines can process these data like humans can do. In order to achieve this aim, ontologies are expected to be used to provide structured vocabularies that describe the relationships between different concepts, allowing computers (and humans) to interpret their meaning in a flexible way and unambiguously. Although there are several ontological languages, OWL [18] is the de facto SW standard ontology language.

2. Semantic Web (SW).

Most of the techniques and inference engines developed for SW data are focusing either on reasoning over instances of an ontology with rules support (e.g. rule-based approaches) or on reasoning over ontology schemas (DL reasoning). Reasoning over instances of an ontology, for example, can derive a certain value for an attribute applied to an object, while reasoning over concepts of an ontology can automatically derive the correct hierarchical location of a new concept in a given concept hierarchy. Nowadays, the integration of rule and DL-based reasoning approaches has also gained a lot of attention and several ontology reasoning systems are currently available, including non-licensed versions like Hermit.

3. Multi-agent systems and intelligent agents

On the other hand, the multi-agent systems and intelligent agents area has received increasing attention by researchers since the end of last century and is currently very SW-relevant. An „Agent” could be defined as a computer system situated in some environment and capable to action autonomously in this environment in order to meet its design objectives. Agents having reactivity (i.e. the ability to perceive its environment and respond to changes to it in a timely fashion), pro-activeness (i.e. the ability to exhibit goal-directed behavior by taking the initiative), and social ability (i.e. the ability to interact with other agents) have been called as the weak notion of agency. Intelligent agents can exhibit some other properties such as temporal continuity (i.e. an agent functions continuously and unceasingly), reasoning (i.e., decision-making mechanism, by which an agent decides to act on the basis of the information it receives, and in accordance with its own objectives to achieve its goals), rationality (i.e. an agent’s mental property that attract it to maximize its achievement and to try to achieve its goals successfully), veracity (i.e. mental property that prevents an agent from knowingly communicating false information), mobility (i.e. the ability for a software agent to migrate from one machine to another), etc.

4. Learning ability of an intelligent agent and of a multi-agent system (MAS)

In particular, one main characteristic of an agent is the learning ability, that is, the capacity to adapt or modify its behavior based on learning experiences. Agents can be useful as standalone entities that are delegated particular tasks on behalf of a user. However, in the majority of cases, agents exist in environments that contain other agents, constituting Multi-agent Systems (MASs). MAS can be seen as a group of agents that can potentially interact with each other. MASs present several advantages over isolated agents, such as reliability and robustness, modularity and scalability, adaptively, concurrency and parallelism, and dynamism.

5. Standardization and integration of agent technology with semantic web services

Efforts toward the standardization of agent technologies have been taken. Organizations such as FIPA (<http://www.fipa.org/>) and OMG Agent PSIG (<http://agent.omg.org/>) are leading this process. In particular, FIPA has become an IEEE Computer Society standards organization aimed at producing standards for the interoperation of heterogeneous software agents FIPA has developed some specifications with a group of normative rules that permit an agent society to operate among themselves. This model identifies some necessary agent’s roles for the platform and agent management: the AMS (Agent Management System) and the DF (Directory Facilitator), which should act as white and yellow pages respectively, and the MTS (Message Transport System), which manages the

interoperability among agent platforms. There exist different FIPA compliant agent platform implementations, like FIPA-Open Source, JADE and ZEUS are the most popular. The agent community is facing the problem of integrating agent technology with Semantic Web Services.

6. Our research about the agent platform required by our VRE.

We are doing research in defining the features of an agent platform organization, tailored to the needs of the problem. It is including flexibility and adaptation to changes as imposed by the VRE management-related knowledge available in the implementation in each moment of time. The agents will have to deal also with various ontologies, due to their evolution in time. Learning should also be a fundamental capability as a way to keep track of the changes in VRE users preferences [22]. Argumentation has been gaining increasing importance, mainly as a vehicle for facilitating rationally justifiable decision making when handling incomplete and potentially inconsistent information.

As the Web grows in size and diversity, there is an increased need to automate aspects of Web Services such as discovery, execution, selection, composition and interoperation.

Composition comprises both choreography, which concerns the interactions of services with their users, and orchestration, which defines the sequence and conditions in which one Web Service invokes other Web Services in order to realize some useful function.

The problem is that current technology around UDDI, WSDL and SOAP provide limited support for all that.

7. Intelligent Web Services thanks to Semantic Web and Web Services.

The joint application of Semantic Web and Web Services in order to create intelligent Web Services is referred to as Semantic Web Services (SWS). SWS consists of describing Web Services with semantic content so that service discovery, composition and invocation can be done automatically. The W3C has examined various approaches with the purpose of reaching a standard for the Semantic Web Services technology, including OWL-S, WSMO, SWSF, WSDL-S, and the proposed as W3C recommendation, SAWSDL. The first three approaches propose an ontology that semantically describes all relevant aspects of Web Services. On the other hand, WSDL-S and SAWSDL identify some WSDL and XML Schema extension attributes that support the semantic description of WSDL components. (OWL) Ontologies, agents and SWS will constitute one of the central pillars of the technological research and development activities to be carried

IV. E-LEARNING WITH THE VRE

E-tutor, supporting learners of an e-learning course, is an alternative concept to the traditional tutoring system. The course tutor in a software tutoring system controls learners relatively weaker than in the traditional one where it is the tutor who is in charge of the support of learning content and fulfilling the assignments. Therefore, in order to obtain better tutoring outcomes, a software tutoring system should emphasize engaging students in the learning process and be adaptive to each individual learner. E-learning offers new possibilities for the learner. The learner can get immediate feedback on his solved problems, can have individualized learning paths, etc.

E-learning services business is growing. The number of organizations working on E-learning and delivering e-learning tools with varying functionality is growing. The number of e-learning courses on the Internet is increasing rapidly [23].

A. Ontologies in E-learning

A structured information representing is required and ontologies (machine process representation containing the semantic information of a domain) can be very useful. The ontology systems serve as a flexible and extendable platform for e-learning management. The inspiring idea to develop reusable atomic learning components and to capture their characteristics in widely-accepted, formal metadata descriptions will most probably attract learning object providers to annotate their products with the accepted standards. An important component of e-learning is testing of student's qualification, skills and knowledge.

For example, in [24] the expediency of computer ontologies use as a transparency tool of European and national qualification frameworks is reasoned. Qualifications are described by triads of professional qualities – knowledge, skills and competencies. A model oriented training helps to compare qualifications and simplifies the procedure for their acceptance. Tools facilitating the correlation of European and national qualification frameworks levels are proposed.

One of the main problems arising during creation of testing systems is an interoperability of created tests – opportunity to reuse these tests in different testing systems. To organize test exchange between various systems it is necessary to create some universal format of tests preservation and their processing instructions. And an important condition for this format should be its independence from the platform. Standardization of educational technologies and, in particular formats of test data preservation is working out all over the world. Now Ministry of Education and Science of Ukraine realize the Program of On-line Education Development.

According to these activities the development of projects of standards for systems, methods and technologies standards of on-line education in educational institutions taking into account international standards was provided. But different test formats such as Instructional Management Systems (IMS) Question

and Test Interoperability (QTI) of Global Learning Consortium are not adequate for the representation of all domain relations.

The more serious problems are caused by the semantic testing. Many authors use the ontology's semantic data to improve the analyses of information in unstructured documents. The domain ontology plays a central role in resource structuring of the learning content. One of the key challenges of the course construction process is to identify the abstract domain of information within which this course will exist. The tutor has to describe the main terms and concepts from which a course is to be constructed.

B. Domain ontology an object of evaluation

The main idea of our approach is that the *domain ontology* is not only the instrument of learning but an object of evaluation of students. We propose for students to build the domain ontology of the study domain and then compare it with the reference one. Results of this comparison show the parts of the domain knowledge which were wrong understood by the student and will help the tutor to improve the e-learning course. Realized experiments demonstrate that this approach is much more efficient than usual tests where some mistakes can be involved by ambiguous formulation of questions and misprints, but correct answers can be obtained intuitively or by accident and don't reflect the real student understanding of the concept about the domain.

Ontological analysis is accomplished by examining the vocabulary that is used to discuss the characteristic objects and processes that are composing the domain, that are developing rigorous definitions of the basic terms in that vocabulary, and that are characterizing the logical connections among those terms. The product of this analysis, *an ontology*, is a domain vocabulary completed with a set of precise definitions, that constrain the meanings of the terms sufficiently to enable consistent interpretation of the data that use that vocabulary [25].

An ontology includes a catalog of terms used in a domain, the rules governing how those terms can be combined to make valid statements about situations in that domain, and the sanctioned inferences that can be made when such statements are used in that domain. In the context of ontology, a relation is a definite descriptor referring to an association in the real world and a term is a definite descriptor that refers to an object or situation-like thing in the real world.

Formal model of ontology O is ordered triple of finite sets $O = \langle T, R, F \rangle$ [15], where T - the domain terms of which is described by ontology O; R - finite set of the relations between terms of domain; F - the domain interpretation functions on the terms and the relations of ontology O. In the process of ontology building, students use relations from the fixed set that contains the most widely used relations: $R = \{ \text{"is a subclass of"}, \text{"is a part of"}, \text{"is a synonym"}, \text{"has attributes"}, \text{"has elements"} \}$. It simplifies the ontology building and analyses processes [26].

The students (as well as the tutor) have to execute four main steps to design the ontology of domain:

1. define the main classes and terms of the domain and describe their meaning: the set of class names T; the set of relation names R;

For every class name define the set of attribute names At; for every attribute name $a \in A_t, t \in T$ define its type - INT, STRING, NUMBER etc. or other class of ontology;

2. Construct the taxonomy of domain terms:

$\langle t_1, t_2 \rangle, t_1 \in T, t_2 \in T, r(t_1, t_2) \rightarrow t_1 \text{ "IS_A_Subclass_Of"} t_2, r \in R$;

3. Define synonymy and other relations:

$\langle t_1, t_2 \rangle, t_1 \in T, t_2 \in T, r(t_1, t_2) \rightarrow t_1 \text{ "IS_Synonyme_Of"} t_2, r \in R$;

$\langle t_1, t_2 \rangle, t_1 \in T, t_2 \in T, r(t_1, t_2) \rightarrow t_1 \text{ "Related_With"} t_2, r \in R$;

4. Describe the instances of constructed classes $\forall a \in t, t \in T$.

We compare the student ontology Os with reference ontology Oe made by tutor:

1. Define the sets of ontology terms Ts and Te;

2. Classify terms from Ts on three disjoint categories: Tn, Tu and Tw. $T_s = T_n \cup T_u \cup T_w$ where correctly defined terms $T_n \subseteq T_e$; not accurately defined terms $T_u \not\subseteq T_e$ but $\forall t_i \in T_n \exists t_{j_1} \in T_e, \dots, t_{j_m} \in T_e, t_{j_k} \in T_e, m = \overline{1, k}$, and incorrectly defined terms $T_w \not\subseteq T_e$ and $\forall t_i \notin T_n \neg \exists t_j \in T_e$;

3. Define the sets of ontology relations Rs and Re;

4. Classify relations from Rs on three disjoint categories: Rn, Ru and Rw. $R_s = R_n \cup R_u \cup R_w$ where correctly defined terms $R_n \subseteq R_e$, not accurately defined terms $R_u \not\subseteq R_e$ but $\forall r_i \in R_n \exists r_{j_1} \in R_e, \dots, r_{j_m} \in R_e, r_{j_k} \in R_e, m = \overline{1, k}$, and incorrectly defined terms $R_w \not\subseteq R_e$ and $\forall r_i \notin R_n \neg \exists r_j \in R_e$;

5. Analyze the use of ontology terms and relations.

We don't consider the use of terms from Tw and relations from Rw. It's very important to take into account the type of relations - hierarchical or improper: Mistake of use "is a part" relation instead of "is a subclass" is much less principle than use "is a sinonime" relation instead of "is a subclass" one.

C. The implementation of the prototype.

Ontological representation of student domain skills can be automatically processed by intelligent software agents. It is appropriate to use software agents for e-learning because they work efficiently in dynamic heterogeneous distributed environment. One of the main properties of an intelligent agent is sociability. Agents are able to communicate between themselves, using some kind of agent communication language, in order to exchange any kind of information. In that way they can engage in complex dialogues, in which they can negotiate, coordinate their actions and collaborate in the

solution of a problem. A set of agents that communicate among themselves to solve problems by using cooperation, coordination and negotiation techniques compose a multi-agent system (MAS). A lot of researchers use MAS for e-learning and e-coaching tasks [27].

M(e)L prototype is a multi-agent ontology-based e-learning system that produces automatic semantic control of student learnt course beliefs. The focus of ontology analysis is on knowledge structuring (of main domain terms and their relations). We use ontologies to describe learning materials and to represent student's belief about the course domain (fig.2).

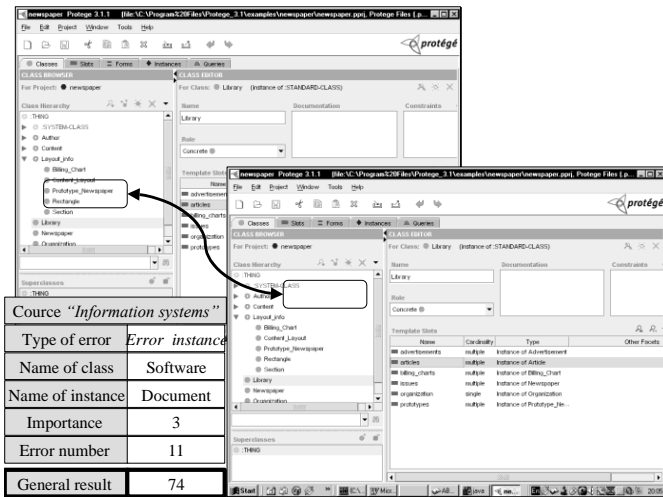


Fig. 2 Domain ontology matching with reference one in M(e)L

V. SUMMARY AND CONCLUSION

A prototype was developed to replace the human tutor intervention. Ontological representation of student domain skills can be automatically processed by intelligent software agents.

The main features of our approach to knowledge control are the following: all results are analysed automatically without human tutor, the results are analysed objectively, students can work with knowledge base, a structuring of domain knowledge simplifies the learning process and tutors can exchange their knowledge based on reference ontologies.

VI. ACKNOWLEDGMENT

This work was supported in part by the project "Design of intelligent system of informational and cognitive support of National qualification frame functioning" of Melitopol State Pedagogic University. We acknowledge of thanks to Rodrigo Martínez-Béjar for proposed information and interesting discussion about perspectives of the Virtual Research Environments.

VII. REFERENCES

- [1] V.Munteanu, C.Sandru, D.Petcu "Multi-cloud resource management: cloud service interfacing", in Journal of Cloud Computing: Advances, Systems and Applications, 3(3), 2014.
- [2] S.Panica, M.Neagul, C.Craciun et al. "Serving Legacy Distributed Applications by a Self-configuring Cloud Processing Platform", Proc. of the 6th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS'2011) I: 2011, pp.139-145.
- [3] M.W.Wallin, G.von Krogh "Organizing for open innovation: focus on the integration of knowledge, Organizational Dynamics", Vol.39 (2), 2010, pp.145-154.
- [4] A.Zuiderwijk, K.Jeffery, M.Janssen. "The necessity of metadata for linked open data and its contribution to policy analyses", in Proceedings of the International Conference for E-Democracy and Open Government CeDEM12, 2012, pp. 281-294.
- [5] R.Studer, R.Benjamins, D.Fensel "Knowledge Engineering", Principles and Methods, Data and Knowledge Engineering, 25(1-2), 1998, pp.161-197.
- [6] J.T.Fernández-Breis, R.Martínez-Béjar "A cooperative framework for integrating ontologies", International Journal of Human-Computer Studies 56(6): 2002, pp.665-720.
- [7] T.R.Gruber. "Nature, nurture, and knowledge acquisition", International Journal of Human-Computer Studies, 71(2): 2013, pp.191-194.
- [8] T.Berners-Lee, J.Hendler "Publishing on the Semantic Web", Nature 410: 2001, pp.1023-1024.
- [9] G.S.Hornby, T.Kurtoglu "Toward a smarter Web", Science 325(5938), 2009, pp.277-278.
- [10] M.Wooldridge. An introduction to MultiAgent Systems. Ed. John Wiley & Sons Ltd, 2002.
- [11] A.di Benedetto A. "Comment on „Is open innovation a field of study or a communication barrier to theory development?", Technovation 30: 557, 2010.
- [12] S.Ontanon, E.Plaza "Learning and joint deliberation through argumentation in multi-agent systems", Proc.of the 6th International Joint Conference on Autonomous Agents and Multi-Agent Systems AAMAS'07, Honolulu, Hawaii, USA, 2007, pp.971-978.
- [13] T.Berners-Lee, J.Hendler, O.Lassila "The Semantic Web", Scientific American, May 2001, pp.34-43.
- [14] W3C (2014). Data Catalog Vocabulary (DCAT). W3C Recommendation 16 January 2014, in Fadi Maali; John Erickson (eds), <http://www.w3.org/TR/2014/REC-vocab-dcat-20140116/> [Accessed 2014-12-23] .
- [15] M. van der Graaf, L.Waaijers "A surfboard for riding the wave: Towards a four country action programme on research data", 2011. – <http://www.knowledge-exchange.info/Default.aspx?ID=469>.
- [16] Web Ontology Working Group OWL Web Ontology Language Guide, 2004. – <http://www.w3.org/TR/owl-guide/>.
- [17] D.Fensel, C.Bussler "The web service modeling framework WSMF", Electronic Commerce Res Appl 1(2), 2002, pp.113-37.
- [18] D.Martin et al. "OWL Web Ontology Language for Services". OWL-S W3C Submission, 2004. – <http://www.w3.org/Submission/OWL-S/>.
- [19] H.Lausen, A.Polleres, D.Roman "Web service modeling ontology", WSMO W3C submission, 2005. – <http://www.w3.org/Submission/WSMO/>.
- [20] S.Battle et al. "Semantic web service framework", SWSF W3C submission, 2005. – <http://www.w3.org/Submission/SWSF/>.
- [21] N.Shadbolt, T.Berners-Lee "Web Science emerges", Scientific American 299 (4), 2008, pp.76-81.
- [22] S.Ontanon., E.Plaza "Learning and joint deliberation through argumentation in multi-agent systems", Proc.of the 6th International Joint Conference on Autonomous Agents and Multi-Agent Systems AAMAS'07, Honolulu, Hawaii, USA, 2007, pp.971-978.
- [23] T. Murray, S. Blessing, S. Ainsworth "Authoring tools for advanced technology learning environments: towards cost-effective adaptive,

interactive, and intelligent educational software”. -
<http://helios.hampshire.edu/~tjmCCS/atoolsbook/chaptersV2/ChapterList.html>.

- [24] S.N.Pryima, A.V. Panin “Use of computer ontologies as a tool to ensure transparency of the European and National qualification frameworks”, *Education technologies and society*, 2013, V.16, no.3, pp. 450–464, <http://ifets.ieee.org/russian/periodical/journal.html>. (in Russian).
- [25] A.Gladun, J.Rogushina, F.Garcia, R.Martínez-Béjar, J.T.Fernández-Breis “An application of intelligent techniques and Semantic Web technologies in e-learning environments”, *Expert Systems with Applications, An International Journal*, 2009, V.36, pp.1922-1931.
- [26] A.Gladun, J.Rogushina, J.Schreurs “Domain Ontology, an Instrument of Semantic Web Knowledge Management in e-Learning”, *International Journal of Advanced Corporate Learning*, V. 5, Issue 4, 2012, pp.21-31.
- [27] A.Gladun, J.Rogushina, V.Shtonda “Ontological Approach to Domain Knowledge Representation for Informational Retrieval in Multiagent Systems”, *Information Theories and Applications*, V.13, N.4, 2006, pp.354-362.

Particle Swarm Optimization Based Discrete Cosine Transform for Person Identification by Gait Recognition

Shahlla A. AbdAlKader

Dept. of Computer Systems, Foundation of Technical Education
Technical Institute, Mosul, Iraq
shahla_ak71@yahoo.com

Omaima N. Ahmad AL-Allaf

Dept. of Basic Sciences,
Faculty of Sciences and Information Technology
AL-Zaytoonah University of Jordan, P.O. Box 130, Amman (11733), Jordan
omaimaalallaf@zuj.edu.jo

Abstract— Gait recognition addresses the problem of human identification at a distance by identifying people based on the way they walk. Therefore, gait recognition has gained growing interest from researchers in recent years. This work presents gait recognition system based on particle swarm optimization (PSO) to recognize a person performing the movement for person identification. The system is based on Discrete Cosine Transform (DCT) for reducing dimensionality and feature extraction. Many experiments were conducted using different: swarm size, block dimension and number of iterations. The results showed that increasing the swarm size to 40 particles and also increasing block size of DCT sub image to (70×70) pixels will increase the overall performance of gait recognition system. The recognition rate reached 96%, MSE reached 0.0088 and PSNR reached 35%.

Keywords— Gait Recognition, Person Identification, Practical Swarm Optimization (PSO), Discrete Cosine Transform (DCT)

I. INTRODUCTION

Person identification can associate an identity to any person. Recently, person identification is highly researched according to its applications such as: authentication to computer systems, buildings, cellular phones and ATMs [1].

Person identification includes many techniques such as token-based, knowledge-based, and biometric-based. Knowledge-based technique depends on something a person knows for identification like password or personal identification number (PIN). Token-based technique depends on something a person has for identification (passport, driver's license, ID card, keys, or credit card). The disadvantages of the two approaches are: tokens may be stolen, lost, forgotten or misplaced. The biometric technique uses physiological or behavioral features of person for identification and it cannot be lost [2]. Fingerprint recognition, iris recognition, face recognition, speech recognition are some of biometric-based techniques [3]. For person identification, these techniques require controlled environment and the person should stand at a standard distance in front of a camera. Therefore, these techniques cannot be used in automatic surveillance of people in real time situations. For this reason, gait recognition has been widely used to provide noninvasive way to recognize persons at a distance without requiring the awareness of the

identified person. Many researches based on gait recognition methods have been proposed in the last decade [1].

Gait or motion can be defined as a sequence of the following poses that recognize people as well as walking. Kinematic chain is a typical representation of a single pose. It describes the pose by a skeleton tree like structure with measured bones lengths. Gait can be captured by a stereovision system of two-dimensional video cameras of typical monitoring systems. Such an acquisition stores motion data in the form of video clips - sequences of the two-dimensional images. There is no direct information about the actor positions, skeleton model and its kinematic chain. Motion capture systems, which acquire motion as a time sequence of poses are much more detailed and accurate [4].

We define gait to be the coordinated, cyclic combination of movements that result in human locomotion. The movements are coordinated in the sense that they must occur with a specific temporal pattern for the gait to occur. The movements in a gait repeat as a walker cycles between steps with alternating feet. It is both the coordinated and cyclic nature of the motion that makes gait a unique phenomenon. Examples of motion that are gaits include walking, running, jogging, and climbing stairs. Sitting down, picking up an object, and throwing and object are all coordinated motions, but they are

not cyclic. Jumping jacks are coordinated and cyclic, but do not result in locomotion [5].

Gait recognition is the process of recognizing many salient properties such as: identity, style of walk, or pathology. This is done based on coordinated and cyclic motions that result in human locomotion [5].

With the increasing demands of visual surveillance systems, human identification at a distance has recently gained more interest. Gait is a potential behavioral feature and many allied studies have demonstrated that it has a rich potential as a biometric for recognition. The development of computer vision techniques has also assured that vision based automatic gait analysis can be gradually achieved. The combination of human motion analysis and biometrics in surveillance systems has become a popular research direction over the past few years. Vision-based human identification at a distance, in particular, has recently gained wider interest from the computer vision community. This interest is strongly driven by the need for automated person identification systems for visual surveillance and monitoring applications in security-sensitive environments such as banks, parking lots, and airports[6][7]. Recently, many researches were focused on gait recognition each with different approaches, advantages and limitations [8..15].

Particle swarm optimization (PSO) is a heuristic, population-based, self-adaptive search optimization technique that is based on swarm intelligence to solve optimization problems in many applications. It comes from the research on the bird and fish flock movement behavior. The algorithm is widely used and rapidly developed for its easy implementation and few particles required to be tuned [16]. PSO was first introduced in 1995 by Kennedy and Eberhart [17] and has been growing rapidly. Many literature researches were focused on developing and enhancing the PSO [18..26]. PSO was used in many researches for solving recognizing problems such as face recognition [27][28] and palmprint recognition [29..32]. We noted that there is lack of literature researches related to gait recognition that based on PSO. Ivekovic et al. (2008) [33] presented PSO for just upper-body pose estimation. They addressed human body pose estimation from still images. They acquired a multi view set of images of a person sitting at a table is and they estimated pose.

Discrete Cosine Transform (DCT) had been introduced by Ahmed, Natarajan and Rao (1974) [34] and can be regarded as a popular transformation technique widely used in image processing [35][36] and. DCT had been used by many researches [35-41] as a feature extraction in recognition process for dimension reduction.

According to above introduction, PSO is used in this work for gait recognition according to its optimization features. To increase the performance of this suggested gait recognition system, DCT will be used for dimensionality reduction and feature extraction. This paper is organized as follows: section

2 includes description of PSO. Section 3 includes description of DCT. Section 4 includes research methodology and section 5 includes results. Finally section 6 concludes this work.

II. PRACTICAL SWARM OPTIMIZATION ALGORITHM

PSO is proposed by Kennedy and Eberhart in 1995 [17]. PSO can be implemented easily, converged rapidly and applied on large number of samples. The PSO includes the following main points [16-26]:

- Each solution is implemented as a particle (N-dimension vector) that represents one individual of a population.
- Each particle has a fitness function (value) associated with it. Each particle adjusts its position and evaluate their position and move closer to optimal point.
- Particles compare themselves to their neighbors and imitate the best of that neighbor.
- Pbest: represents the best value of the particle i.
- Gbest: best value that one of the swarm particle reach it.
- Lbest: best value that particle in a local swarm reach it
- Eq.1 used to compute new velocity of each particle:

$$V_i(t+1) = W \times V_i(t) + C1 \times \text{rand} \times (Pbest(t) - X_i(t)) + C2 \times \text{rand} \times (Gbest(t) - X_i(t)) \dots\dots\dots(1)$$

Where, V[]: particle velocity,
 Xi: ith particle of swarm
 W: weight (random number between 0 and 1).
 C1, C2 : the speeding factors (with value 2).
 From Eq.1, the new velocity vi(t+1) is affected by: Pbest, Gbest and Vi(t): velocity of ith particle X in time t.

- Eq.2 used to compute new fitness value of each particle:

$$X_i(t+1) = X_i(t) + V_i(t+1) \dots\dots\dots(2)$$

 The particle will change its value according to its new velocity (vi(t+1))

PSO algorithm was described in details in researches [16-26]:

1. Initialize parameters (number of generations, population size, weights, c1, c2)
2. Initialize population (velocity and position of each particle) and initialize Pbest and Gbest.
3. New generation
4. Take one particle(P) from population
5. Compute new velocity (Pvelocity) of particle using Eq.1.
6. Compute new position (Pposition) of particle using Eq.2
7. Pbest = Pposition if cost(Pposition <= cost(Pbest)
8. Gbest = Pbest if cost(Pbest) <=cost(Gbest)
9. Repeat steps (4..10) until there are more particles in population
10. Repeat steps (3..10) until reaching maximum number of generations
11. Return Gbest

III. DISCRETE COSINE TRANSFORM (DCT)

DCT transforms the input image into a linear combination of weighted basis functions. DCT transform image from spatial domain to frequency domain. DCT uses cosine base functions and exhibits good de correlation and energy compaction

$$F(u, v) = \alpha(u)\alpha(v) \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} \cdot \cos\left[\frac{\pi \cdot u}{2 \cdot N} (2x + 1)\right] \cos\left[\frac{\pi \cdot v}{2 \cdot M} (2y + 1)\right] f(x, y) \dots\dots\dots(3)$$

Where f (x, y) is the intensity of pixel in row x and column y,
 u = 0,1, ..., N-1,
 v =0,1,..., M-1,
 α(u), α(v): functions are defined as following equation [34-36]:

$$\alpha(u), \alpha(v) = \begin{cases} \sqrt{\frac{1}{N}} & \text{for } u,v =0 \\ \sqrt{\frac{2}{N}} & \text{for } u,v \neq 0 \end{cases} \dots\dots\dots(4)$$

The DCT helps separate the image into parts of differing importance with respect to image's visual quality. For most images, much of signal energy lies at low frequencies. These are relocated to upper-left corner of DCT array. Lower-right values of DCT array represent higher frequencies and turn out to be small to be removed with little visible distortion. The number of DCT coefficients might affect the recognition rate. DCT had been used by many researches as a feature reduction and extraction [34-36].

IV. RESEARCH METHODOLOGY

The research methodology depends on a database taken from CASIA [42] database with different views that have different silhouette in person's height and width. The Institute of Automation, Chinese Academy of Sciences (CASIA) provide the CASIA Gait Database to gait recognition and related researchers to promote the research. In CASIA Gait Database there are three datasets: Dataset A, Dataset B (multi view dataset) and Dataset C (infrared dataset). Dataset A (former NLPR Gait Database) was created on 10Dec2001 including 20 persons.

Each person has 12 image sequences, 4 sequences for each of three directions (parallel, 45° and 90° to image plane). The length of each sequence is not identical for the variation of the walker's speed, but it must ranges from 37 to 127.

A gait recognition system based on PSO and DCT for feature extraction is suggested in this work. The main steps for PSO for training/testing gait recognition system were implemented using MatLab2013. The DataBase of this system includes 9000 images (each of size 240×352 pixels) of 15 persons which selected from CASIA database. Each person with 50 images (states) for 4 cases for three angles (0, 45 and 90). At the end the selected database includes: 15 person × 3 angles × 4 cases × 50 states =9000 images.

characteristics. DCT of an N×M image f(x, y) is defined by the following equation [34-36]:

A. Training Part Of Gait Recognition

The training part of the gait recognition system is described in Fig.1 and includes the following steps:

1. Read 50 images (each of size 240×352 pixels) for each one of the 4 states for each one of the three angles (0, 45 and 90).
2. The OR logical gate will be applied on each 50 images to produce only one average image for each case of the 4 cases. This is applied for each angle. Then the total number of images resulted from this process are: 15 × 3 × 4 × 1= 180 images for 15 persons. Fig.2 shows the image of person 1 after applying the OR gate on 50 images of gait of person1 for angle 90°. Fig.3 shows the image of person 1 after applying the OR gate on 50 images of gait of person1 for angle 0°. Whereas Fig.4 shows the image of person 2 after applying the OR gate on 50 images of gait of person1 for angle 45°.
3. Data standardization. Resize each one of image of size 240×352 pixels to be image of size 190×100 pixels. The main goal is producing a dataset with the same position of the person in the middle of each frame and same size in whole image sequence. The idea is to fix the head for each frame in a predefined position and resize the body to achieve a preset height. We perform a three stage preprocessing: extract rectangle including the person without extra black pixels and obtain height and width of the person; sequence is calculated and each frame is converted to biggest height and width; and finally, move head of each frame in a fixed point.
4. Take small block (70×70, 60×60, 40×40 or 20×20) from each image of size 190×100 pixels. We will use different dimension for each experiment to examine which dimension will lead to best recognition rate.
5. Convert each sub image block from two dimensional array to one dimensional array.
6. The person properties will be extracted by applying DCT algorithm for feature extraction.
7. PSO for classification is used for each one of the 180 feature vectors (generated using DCT) as follows:

- Step1: initialize PSO parameters as shown in Table I.
- Step 2: initialize position, velocity, Pbest and Gbest.
- Step3: Calculate fitness function of each sample
- Step4: Calculate optimal value of particle swarm (Pbest) and optimum value of group (Gbest) according to comparison between current value of particle and Pbest and Gbest
- Step5: Calculate new speed of practical according to Eq.1.
- Step6: Compute new position of particle according to Eq.2.
- Step7: Repeat steps (3-6) while more iterations.
- Step8: Store features sub set which are represented by vector with 40 values (according to population size) in sub features database: gaitdbf

C1	2
C2	2
Weight	0.5
Number of Iterations	100,150

B. Testing Part of Gait Recognition

The testing part (recognition step) of the suggested system is described in details in Fig.5.

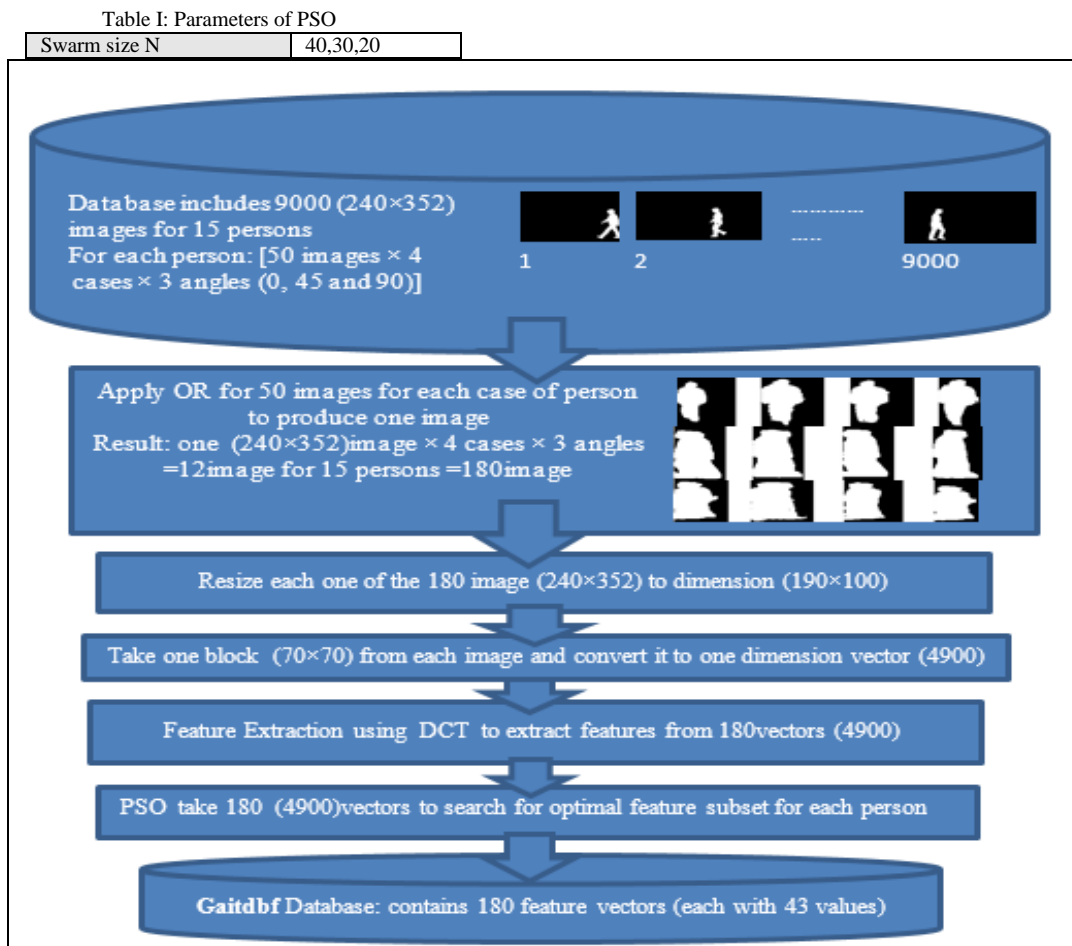


Fig.1: The Training Part of gait recognition system using PSO and DCT



Fig.2: Image of person 1 after applying OR on 50 images for each state of gait of person1 for angle 90°



Fig.3: Image of person 1 after applying OR on 50 images of gait of person1 for angle 0⁰

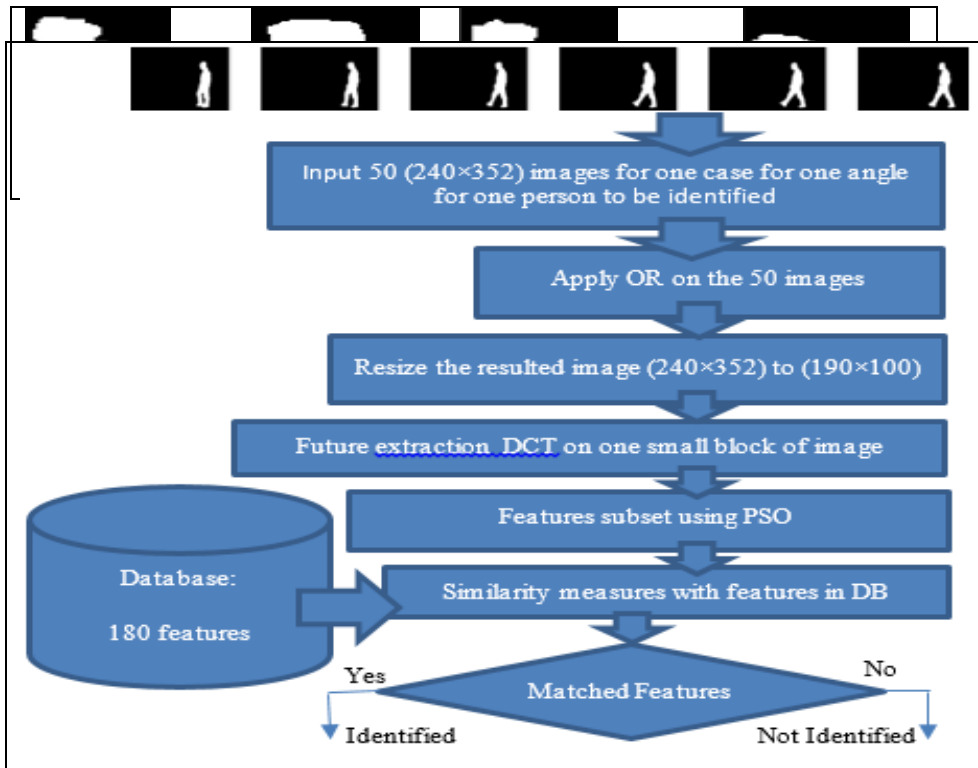


Fig.5: The Testing Part of gait recognition system using PSO and DCT

V. EXPERIMENTAL RESULTS

The suggested gait recognition system that is based on PSO and DCT is implemented using MATLAB 2013. Many experiments were conducted for the suggested gait recognition system. The DataBase of the suggested gait recognition system includes 9000 images (each of size 240x352 pixels) for 15 persons were selected from CASIA DataBase [42]. Each person with 3 angles (0, 45 and 90), each angle with 4 cases and 50 images for each case. Each original image is resized from its original dimension to 190x100 pixels. The performance of the suggested system is computed using recognition ratio, MSE and PSNR.

The first experiment depends on executing the PSO and DCT for feature extraction using various DCT coefficients sizes. The two dimensional DCT is applied to the input image and only a subset of DCT coefficients corresponding to the upper left corner of DCT array is retained.

Different subset sizes of 70x70, 60x60, 40x40 and 20x20 of the original 100x190 DCT array are used in this experiment as input to the subsequent feature selection phase.

In this experiment, we determined swarm size equal 40 and 100 iterations. Table II shows the results of using PSO/ DCT with different block size.

TABLE II: PSO/DCT (SWARM=40, ITERATIONS=100)

Swarm size N=40, No. of Iterations=100			
subimage	reco.rate	MSE	PSNR
70×70	96%	0.0088	35
60×60	92%	0.0121	32
40×40	89%	0.0187	29
20×20	85%	0.0211	27

We can note from Table II that best results including high recognition rate and low MSE were obtained when selecting sub image dimension equal 70×70.

Another experiment was conducted for PSO/DCT with selection of swarm size equal 40 and number of iterations equal 150. At the same time, we determined different sub image dimension for this experiment (70×70, 60×60, 40×40 or 20×20). Table III shows results of PSO/DCT with swarm size N=40 and number of iterations equal 150.

TABLE III: PSO/DCT RESULTS (SWARM=40,ITERATIONS=150)

Swarm size N=40 and No. of Iteration=150			
subsetimage	reco.rate	MSE	PSNR
70×70	95%	0.0089	34
60×60	91%	0.0134	30
40×40	87%	0.0178	28
20×20	84%	0.0256	26

Other experiments were based on PSO/DCT with swarm size equal 30, number of iterations equal 100 and using different size of sub image dimension (70×70, 60×60, 40×40 or 20×20). Table IV shows results of PSO/DCT with swarm size N=30 and number of iterations equal 100.

TABLE IV: PSO/DCT RESULTS (SWARM=30, ITERATIONS=100)

Swarm size N=30 AND No. of Iteration=100			
Subset image	Reco.rate	MSE	PSNR
70×70	95%	0.0091	34
60×60	91%	0.0125	31
40×40	88%	0.0189	28
20×20	84%	0.0217	27

Other experiments were based on PSO/DCT with swarm size equal 30, number of iterations equal 150, with different size of sub image dimension (70×70, 60×60, 40×40 or 20×20). Table V shows results of PSO based LDA/DCT with Swarm size N=30 and number of iteration=150

TABLE V: PSO/DCT RESULTS (SWARM=30, ITERATION=150)

Swarm size N=30 AND No. of Iteration=150			
Subset image	Reco.rate	MSE	PSNR
70×70	91%	0.0119	32
60×60	87%	0.0178	28
40×40	86%	0.0220	27
20×20	84%	0.0276	26

Finally, other experiments were based on PSO/DCT with swarm size equal 20, number of iterations equal 150, with different size of sub image dimension (70×70, 60×60, 40×40

or 20×20). Table VI shows results of PSO based LDA/DCT with Swarm size N=20 and No. of Iteration=150.

TABLE VI: PSO/DCT RESULTS (SWARM=20, ITERATION=150)

Swarm size N=20 and No. of Iteration=150			
Subset image	Reco.rate	MSE	PSNR
70×70	90%	0.0122	30
60×60	86%	0.0187	26
40×40	84%	0.0224	25
20×20	82%	0.0287	24

From Table III, Table IV, Table V, and Table VI, we can note that the swarm size can affect the overall results (recognition rate, MSE and PSNR). The best results of recognition rate were obtained when selecting swarm size equal 40 with 100 iterations. Also the dimension of the sub image can affect the recognition rate of the system. Best recognition rates for all experiments were obtained when determining sub image dimension equal 70×70. This is because the big sub image dimension will take more features of the image. Whereas the results related to recognition rate are low when determining sub image dimension equal 20×20 because small features of image will be taken. Fig.6 shows that increasing the block size will increase the recognition rate and PSNR of the gait recognition system when selecting swarm size equal 40 and number of iterations equal 100. Fig.7 shows the effect of swarm size on the recognition rate and PSNR when selection bloc dimension equal 70×70 and 150 number of iterations

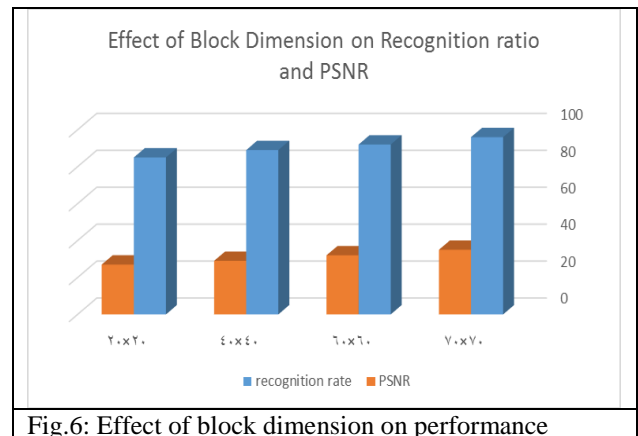


Fig.6: Effect of block dimension on performance

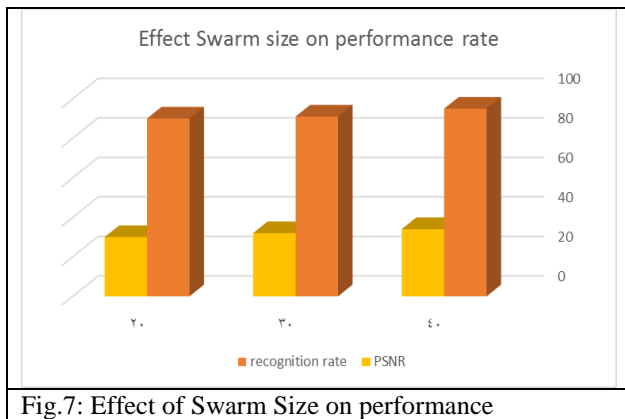


Fig.7: Effect of Swarm Size on performance

VI. CONCLUSION

Gait recognition is a type of biometric recognition and related to the behavioral characteristics of biometric recognition. Person identification using gait is method to identify an individual by the way he walk. A gait recognition system was presented in this paper using PSO and DCT for feature reduction and extraction

The gait recognition system was implemented using MatLab 2013. The DataBase of the gait recognition program includes 9000 images (each of size 240×352 pixels) of 15 persons which selected from CASIA database [42] with different angles (0, 45 and 90), cases (4 cases) and states (50 state for each person). The original images were resized from 240×352 pixels to 190×100 pixels.

Many experiments were conducted for executing the gait recognition program based on PSO and DCT with different: swarm size, number of iterations and sub block dimension. The experimental results showed that the best values of recognition rate, MSE and PSNR were obtained when increasing the sub image block size of 70×70 pixels. Also best results were obtained when increasing the swarm size to 40. The recognition rate reached 96%, MSE reached 0.0088 and finally PSNR reached 35%.

As a future work, other feature extraction algorithm may be used to reduce the image dimensionality and feature extraction of the image to be used in recognition process. Many experiments will be conducted to make comparisons between different algorithms for feature extraction to determine the suitable algorithm that lead to best recognition performance.

REFERENCES

[1] Sruthy Sebastian, Activity Based Person Identification Using Particle Swarm Optimization Algorithm, International Journal of Computer Science and Mobile Computing, Vol.2, Issue.7, Jul2013, pp:1-6.
 [2] A. K. Jain, A. Ross, S. Prabhakar, "An Introduction to Biometric Recognition", IEEE Trans. on Circuits and Systems for Video Technology, Vol.14, No.1, Jan2004, pp 4-19.

[3] Gajanan P. K., et al. Human Computer Interpreting with Biometric Recognition System, International Journal of Advanced Research in Computer Science and Software Engineering, Vol.2, Issue.12, Dec2012, pp:140-147.
 [4] Adam Switoński, et al. Human Identification Based on the Reduced Kinematic Data of the Gait, 7th International Symposium on Image and Signal Processing and Analysis (ISPA), IEEE, Dubrovnik,4-6 Sep2011, pp:650-655.
 [5] M.Tistarelli, J.Bigun, and E.Grosso, Biometric Gait Recognition, Biometrics School 2003, LNCS 3161, 2005, pp: 19-42.
 [6] Liang Wang, et al. Silhouette Analysis-Based Gait Recognition for Human Identification, IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol.25, No.12, Dec2003, pp:1-14.
 [7] Liang Wang, , et al, Fusion of Static and Dynamic Body Biometrics for Gait Recognition, Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003), Vol.2.
 [8] Chiraz BenAbdelkader, et al. EigenGait: Motion-based Recognition of People using Image Self-Similarity, Lecture Notes in Computer Science Vol.2091, 2001, pp:284-294.
 [9] Payam S. and Swarup M. and Yuri O., Multi-View Classifier Swarms for Pedestrian Detection and Tracking, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR Workshops, San Diego, CA, USA, 25-25Jun2005.
 [10] Qiong C., Bo F., and Hui C., Gait Recognition Based on PCA and LDA, Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCCT '09), Huangshan, P. R. China, 26-28Dec2009, pp:124-127.
 [11] Ra'ul M. and Tao Xiang, Gait Recognition by Ranking, A. Fitzgibbon et al. (Eds.): ECCV 2012, Part I, LNCS 7572, Springer-Verlag Berlin Heidelberg,2012, pp:328-341.
 [12] Mohamed Rafi, et al. A Model Based Approach for Gait Recognition System, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Vol.3, Issue.5, Nov 2013, pp:223-228.
 [13] Bogdan Kwolek, et al. 3D Gait Recognition Using Spatio-Temporal Motion Descriptors, DOI: 10.1007/978-3-319-05458-2_61 In book: Intelligent Information and Database Systems, Publisher: Springer International Publishing, Editors: Nguyen, NgocThanh and Attachoo, Boonwat and Trawiński, Bogdan and Somboonviwat, Kulwadee, 2014, pp.595-604.
 [14] Rong Z., Christian V. and Dimitris M., Human Gait Recognition, Conference on Computer Vision and Pattern Recognition Workshop, 27-02Jun2004. CVPRW '04.
 [15] Adam S., Andrzej P., Konrad W., Human Identification Based on Gait Paths, book chapter: Advanced Concepts for Intelligent Vision Systems, 13th International Conference, ACIVS 2011, Ghent, Belgium, 22-25 Aug 2011, pp 531-542.
 [16] Qinghai Bai, Analysis of Particle Swarm Optimization Algorithm, Computer and Information Science, Vol.3, No.1, 2010.
 [17] Kennedy, J. and Eberhart, R., Particle swarm optimization, Proceedings of IEEE International Conference on Neural Networks, Perth, WA, 1995, pp:1942-1948.
 [18] Dian P.R., Siti M.S. and Siti S.Y., Particle Swarm Optimization: Technique, System and Challenges, International Journal of Computer Applications (0975 – 8887) Vol.14, No.1, Jan2011, pp:19-27.
 [19] Satyobroto Talukder, Mathematical Modelling and Applications of Particle Swarm Optimization, Master's Thesis, Mathematical

- Modelling and Simulation, School of Engineering at Blekinge Institute of Technology, Master of Science, Feb 2011.
- [20] Voratas K., Comparison of Three Evolutionary Algorithms: GA, PSO, and DE, *Industrial Engineering & Management Systems*, Vol.11, No.3, Sep2012, pp:215-223.
- [21] H. Kuo, J. Chang and C. Liu, Particle Swarm Optimization For Global Optimization Problems, *Journal of Marine Science and Technology*, Vol. 14, No. 3, 2006, pp:170-181.
- [22] Daniel Bratton and James Kennedy, Defining a Standard for Particle Swarm Optimization, *Proceedings of the 2007 IEEE Swarm Intelligence Symposium, SIS, 2007*.
- [23] Davoud S. and Ellips M., Particle Swarm Optimization Methods, Taxonomy and Applications, *International Journal of Computer Theory and Engineering*, Vol. 1, No. 5, Dec 2009, pp: 486-502.
- [24] M. Peyvandi, M. Zafarani and E. Nasr, Comparison of Particle Swarm Optimization and the Genetic Algorithm in the Improvement of Power System Stability by an SSSC-based Controller, *Journal of Electrical Engineering & Technology* Vol. 6, No. 2, , 2011, pp:182-191.
- [25] Bijayalaxmi Panda, Soumya Sahoo, Sovan Kumar Patnaik, A Comparative Study of Hard and Soft Clustering Using Swarm Optimization, *International Journal of Scientific & Engineering Research*, Vol.4, Issue.10, Oct2013, pp:785-790.
- [26] Rania Hassan, Babak Cohanim, Olivier de Weck, A Copmarison Of Particle Swarm Optimization And The Genetic Algorithm, *American Institute of Aeronautics and Astronautics, 46th Aiaa/Asme/Asce/Ahs/Asc Structures, Structural Dynamics And Materials Conference*, 2005.
- [27] R. M. Ramadan and R. F. AbdelKader, Face Recognition Using Particle Swarm Optimization-Based Selected Features, *International Journal of Signal Processing, Image Processing and Pattern Recognition*, Vol.2, No.2, Jun2009, pp:51-66.
- [28] P.V. Shinde, B.L. Gunjal and R. G. Ghule, Face Recognition Using Particle Swarm Optimization, *Emerging Trends in Computer Science and Information Technology - 2012(ETCSIT2012) Proceedings published in International Journal of Computer Applications (IJCA)*, pp:11-13.
- [29] M. Arunkumar And S. Valarmathy, Palmprint And Face Based Multimodal Recognition Using Pso Dependent Feature Level Fusion, *Journal Of Theoretical And Applied Information Technology*, Vol.57, No.3, 30November 2013, pp:337-346.
- [30] K. Krishneswari And S. Arumugam, Intra Modal Feature Fusion Based On PSO For Palmprint Authentication, *ICTACT Journal On Image And Video Processing*, May 2012, Vol.2, Issue.4, pp:435-440.
- [31] Ola M. Aly, et al., A Multimodal Biometric Recognition system using feature fusion based on PSO, *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 2, Issue.11, Nov2013, pp: 4336-4343.
- [32] Ola M. Aly, et al., An Adaptive Multimodal Biometrics System using PSO, *International Journal of Advanced Computer Science and Applications*, Vol. 4, No.7, 2013, pp:158-165.
- [33] S. Ivekovic, E. Trucco and Y. R. Petillot, Human Body Pose Estimation With Particle Swarm Optimisation, *Evolutionary Computation*, Vol.16, No.4, pp: 509-528.
- [34] M. Ahmad, T. Natarajan and K. R. Rao, Discrete Cosine Transform, *IEEE Trans. Computers*, 1974, pp:90-94.
- [35] Aman R. C., Pallavi P. V. and M. M. Roja, Face Recognition Using Discrete Cosine Transform for Global and Local Features, *Proceedings of the 2011 International Conference on Recent Advancements in Electrical, Electronics and Control Engineering (ICONRAEECE), IEEE Xplore.*, 2011.
- [36] Ziad M. Hafed and Martin D. Levine, Face Recognition Using the Discrete Cosine Transform, *International Journal of Computer Vision*, Vol.43, No.3, 2001 Kluwer Academic Publishers. Manufactured in The Netherlands, 2001, pp:167-188.
- [37] C. Podilchuk and X. Zhang, "Face Recognition Using DCT-Based Feature Vectors," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'96)*, vol. 4, May1996, pp. 2144-2147.
- [38] Z. Yankun and L. Chongqing, "Efficient Face Recognition Method based on DCT and LDA," *Journal of Systems Engineering and Electronics*, Vol. 15, No. 2, pp:211-216, 2004.
- [39] F. M. Matos, L. V. Batista, and J. Poel, "Face Recognition Using DCT Coefficients Selection," *Proc. of the 2008 ACM Symposium on Applied Computing, (SAC'08)*, March 2008, pp:1753-1757.
- [40] Z. Pan and H. Bolouri, "High Speed Face Recognition Based on Discrete Cosine Transform and Neural Networks," *Technical Report, Science and Technology Research Center (STRC), University of Hertfordshire*.
- [41] Virendra P. Vishwakarma, Sujata Pandey Member IEEE, and M. N. Gupta , A Novel Approach for Face Recognition Using DCT Coefficients Re-scaling for Illumination Normalization, *15th International Conference on Advanced Computing and Communications, IEEE Computer Society*, pp:535-539.
- [42] CASIA Gait Database, [http:// www.sinobiometrics.com](http://www.sinobiometrics.com), 2006. CASIA Gait Database collected by Institute of Automation, Chinese Academy of Sciences" and a citation to "CASIA Gait Database, <http://www.sinobiometrics.com>" should be added into the references, 2006.

Supporting Arabic Sign Language Recognition with Facial Expressions

Ghada Dahy Fathy

Faculty of Computers and Information
Cairo University
Cairo, Egypt
g.dahy@fci-cu.edu.eg

E.Emary

Faculty of Computers and Information
Cairo University
Cairo, Egypt
e.emary@fci-cu.edu.eg

Hesham N.ElMahdy

Faculty of Computers and Information
Cairo University
Cairo, Egypt
ehesham@cu.edu.eg

Abstract—this paper presents an automatic translation model for combination of facial expressions of user and gestures of manual alphabets in the Arabic sign language. The part of facial expression depends on locations of user's mouth, nose and eyes. The part of gestures of manual alphabets in the Arabic sign language does not rely on using any gloves or visual markings to accomplish the recognition job. As an alternative, it deals with images of signer's hands. Two parts enable the user to interact with the environment in a natural way. First part in the model deals with signs and consists of three phases preprocessing phase, skin detection phase and feature extraction phase. Second part in the model that deals with facial expressions consists of two phases face detection and tracking facial expression. Proposed model has an accuracy 90% using minimum distance classifier (MDC) and absolute difference classifier in case of facial expressions and 99% in case of signer's hands.

Keywords—Arabic Sign Language, Facial Expression, Minimum Distance Classifier (MDC), Human computer interaction (HCI), Absolute Distance Classifier (ADC).

I. INTRODUCTION

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6]. Signing has always been part of human communications. Sign language (SL) is a form of manual communication and is one of the important communications for people in deaf community [1]. For thousands of years, deaf people have generated and used signs among themselves. In the past, signs were the only ways of communication available for all deaf people. The sign language is the fundamental communication method between people who suffer from hearing impairments. As we know about oral language, sign language is not universal because it has different features that differ from country to other according to the country, or even according to the regions. Sign language in the Arab World has recently been

recognized and documented. Very great efforts have been made to build the sign language used in individual countries, including Jordan, Egypt and the Gulf States, by trying to standardize the language and spread it among members of the deaf environment. In the recent years, the idea of the computerized translator became an interesting research area [2]. There are two ways for interacting between human and computer: glove-based and vision-based systems [3]. The glove-based system depends on electromechanical devices that are used for data collection about the gestures [4]. The user has worn some sort of gloves that cover with sensors to make the interaction between the system and computer. According to readings of sensors signs meaning will be understood. There is difficult for signers in moving with great numbers of sensors so second way of human computer interaction (HCI) has been provided to overcome this

problem. Second way depending on image of signers in their communication use two channels: manual and non manual. In the manual channel, deaf people use their hands to express lexical meaning. In the non manual channel deaf people use their facial expression, upper body movements and head to express syntactic and semantic information. Non-manual expression co-occurs with manual signs to support users. In this work, our goal is to construct a model that is able to translate Arabic sign language (ASL) to Arabic text. We take in our consideration grammatical expressions that provide the grammatical structure of sentence. We use four face emotions for dealing with non-manual expression neutral, sad, happy, and angry. Each type composed of a combination of facial features movements. For identifying facial expression in sign language we tracked sets of features in faces image like eyes, nose and mouth locations. The paper is composed of six main sections. First section will be about related works that discuss previous work in sign language and facial expressions second section will be about proposed schema model and how we extract features of signs and facial expressions. Third section will be about methodology in sign language part. The fourth section will be about methodology in facial expressions part. Experimental results will be discussed in section number five. The last section will contain summary about paper and future work.

II. II. RELATED WORK

In recent years, several research projects in developing sign language systems were presented [5]. An Arabic Sign Language Translation Systems (ArSL-TS) [6] model has been introduced. That model for sign language runs on mobile devices that model enable users to translate Arabic text into Arabic Sign Language for the deaf on mobile devices such as Personal Digital Assistants (PDAs). Software in [7] consists of two basic modules: linguistic translation from printed English into sign language, and virtual human animation. The animation software enables Simon to sign in real-time. A dictionary of signed words makes system to look up the accompanying physical movement, facial expressions and body positions, which are stored as motion-capture data on a hard disk. This model contains very realistic and accurate hand representations, developed within the project. Moreover, natural skin textures are applied to the hands and face of the model to generate the maximum impression of subjective reality. In [8], an automatic Thai finger-spelling sign language translation system was developed using Fuzzy C-Means (FCM) and Scale Invariant Feature Transform (SIFT) algorithms. Key frames took from several subjects at different times of day and for several days. Also, testing Thai finger spelling words video took from 4 subjects with the SIFT threshold of 0.7 and use one nearest neighbor prototype. In [9], an automatic translation of static gestures of alphabets in American Sign Language (ASL) was developed, ASL used three feature

extraction methods and used neural network to classify signs. The proposed system interacts with images of bare hands, which allows user to interact with environment in as normal people. Token image would be processed and converted to a feature vector that will be compared with the feature vectors of a training set of signs. The system is implemented and tested using data sets of hand images samples for each signs. System used three feature extraction methods are tested and the best method is suggested with results obtained from Artificial Neural Network (ANN). Recent works on tracking facial features used sets of Active Shape Models to constrain face shapes and also considered head motions [9], [10]. KLT was used in [11] to track facial feature points, but it had problem because their 2D local models for shape constraints that were based on frontal face might not cope well under varying head pose.

Algorithm 1 Facial Feature Extraction.

- 1: Get frames that contain facial movement.
- 2: Apply median filter with 3×3 windows to remove noise from frames.
- 3: Convert RGB image into YCbCr to detect skin.
- 4: Calculate first component in YCbCr

$$Y = 16 + (65.481.R + 128.553.G + 24.966.B). \quad (1)$$

- 5: Calculate second component in YCbCr

$$Cb = 128 + (-37.797.R - 74.203.G + 112.0B). \quad (2)$$

- 6: Calculate third component in YCbCr

$$Cr = 128 + (112.0.R - 93.786.G + 18.214.B). \quad (3)$$

- 7: Mark skin pixel to detect face that contain $Cb \geq 77, Cb \geq 127, Cr \geq 133$ and $Cr \geq 173$.
- 8: Detect boundaries by using Sobel (15) after that applying horizontal projection to mark eyes region. Taking the upper half of face and calculate the vertical projection to separate eyebrows from eyes.
- 9: Select the lower part of face and calculate vertical projection to get mouth and nose region.
- 10: Draw rectangular box on each of the detected feature elements.
- 11: Generate feature vector of width and height of each rectangular.

Algorithm2 Signs Feature Extraction

- 1: Get video that represent hand movement.
- 2: Divide video into frames.
- 3: Apply median filter with 3×3 windows to remove noise from frames.
- 4: Convert RGB image into YCbCr to detect skin.
- 5: Calculate first component in YCbCr

$$Y = 16 + (65.481.R + 128.553.G + 24.966.B) \quad (4)$$

6: Calculate second component in YCbCr

$$Cb = 128 + (-37.797.R - 74.203.G + 112.0B) \quad (5)$$

7: Calculate third component in YCbCr

$$Cr = 128 + (112.0.R - 93.786.G + 18.214.B) \quad (6)$$

8: Mark skin pixel that contain $Cb \geq 77$, $Cb \geq 127$, $Cr \geq 133$ and $Cr \geq 173$ and crop image that contain skin.

9: Divide cropped image into blocks each block with size 4×4 pixels.

10: Get centroid of block

$$\text{Centroid of block} = \frac{\sum_{i=1}^{16} \text{Element } i \text{ of block}}{16} \quad (7)$$

11: Store all centroids of all blocks as extracted features

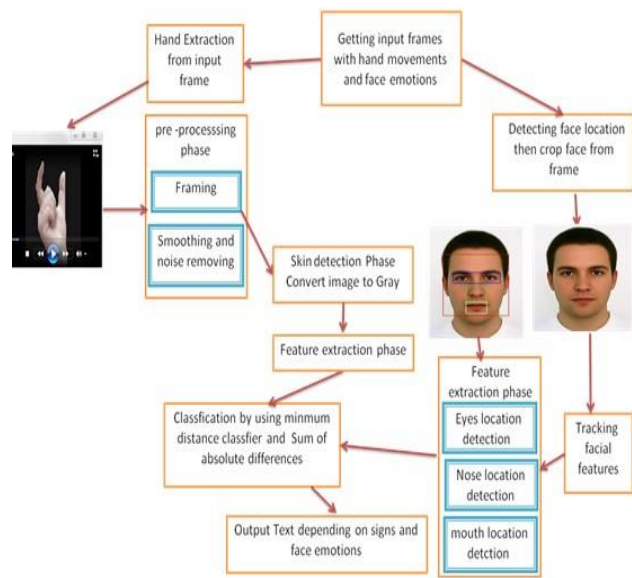


Fig. 1: Facial expressions and Arabic sign language model architecture.

III. THE PROPOSED APPROACH

Facial expressions in sign language model are composed of three main phases for feature extraction of signs namely, Pre-processing phase, Skin detection phase and Feature extraction. Model composed of 2 main phases for facial expressions extraction namely, face detection and tracking facial features. Figure 1 depicts the structure of the Facial expressions in sign language model. Pre-processing phase in the signs part receives, as an input, a video that contains the signed words to be translated into text and prepare it to be ready for use in subsequent phases. Skin detection phase in the signs part detect skin in image by converting RGB image into YCbCr formatting. YCbCr is a family of color spaces. YCbCr has better accuracy compared with other color spaces families in case of skin detection. YCbCr presents color as bright-ness and two color difference signals. Components Y is the brightness (luma), Cb and Cr are two colors

Difference signals. Model calculates YCbCr components by using equations in algorithm 2 after that model converts input frame into gray to enables us in defining and separating location of hands and background. Finally we extract features from input frame. Detecting and cropping face in the second part of the system is very important phase. For face detection we use YCbCr color space model to define the location of face in the image. Values of Cb and Cr component support model in defining the skin part in the input frames as we shown in algorithm 1. In classification phase, each unknown facial expressions or signs are being matched with all the known expressions and signs in the same category in the database and takes the nearest one to expressions. Database of the model deals with 7 facial expressions neutral, smile, sad, angry, afraid, disgusted and surprised. There are 105 samples of facial expressions. 15 samples for neutral face, 15 samples for smile face, 15 samples for sad, 15 sample for angry face, 15 samples for afraid face, 15 samples for disgusted, 15 samples for surprised face. Database of the model contains dictionary for all Arabic signs.

IV. METHODOLOGY: PHASE-I

A. Pre-Processing

Firstly, a video that contains stream of signed words (gestures) to be translated is acquired. After that, the video enters the pre-processing phase where video divides into frames. Then, smoothing is applied for each frame to remove noise by using median filter with 3×3 windows. The median filter considers each pixel in the image in turn and looks at its nearby neighbors to decide whether or not it is representative of its surroundings. The median is calculated by first sorting all the pixel values from the surrounding neighborhood into numerical order and then replacing the pixel being considered with the middle pixel value (If the neighborhood under consideration contains an even number of pixels, the average of the two middle pixel values is used).

B. Skin Detection

In that phase system tries to detect the skin part in the input frame because the skin part represents hands in the frame

Firstly system converts RGB image into YCbCr image. System calculates YCbCr components by using equations in algorithm 2. In figure 2, we see the difference between original image and YCbCr image. Finally system converts the frame into gray by using equation number 8 to isolates skin with black color from background as shown in figure 3.

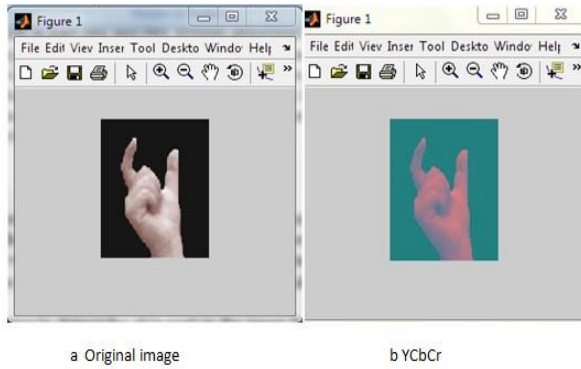


Fig. 2: Converting Original image into YCbCr.
 $Gray(x) = 0.29R + 0.59G + 0.11B.$ (8)

- Where x is the input pixel.
- R is the red value of input pixel.
- G is the green value of input pixel.
- B is the blue value of input pixel.

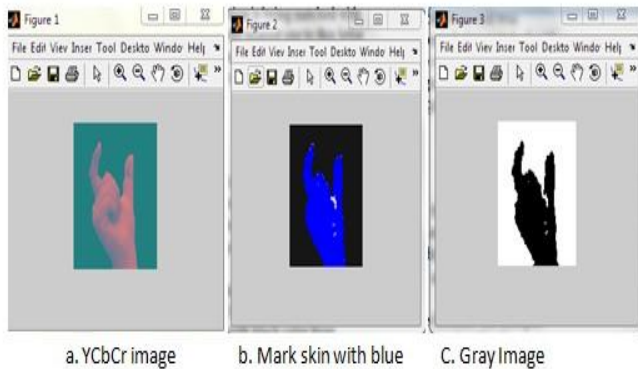


Fig. 3: Skin detection process

C. Feature Extraction

Feature extraction phase depends on Centroid. Firstly system divides the input frame into blocks with size 4X 4. In that model we use centroid properties for extracting features from blocks as we shown in equation 7 in algorithm 2.

V. METHODOLOGY: PHASE-II

A. Detecting and cropping phase

Recognition algorithms divide into two main approaches, geometric, that depends on distinguishing features, or photometric, which is a statistical approach that distills an image into values and compares the values with templates to eliminate variances. Popular recognition algorithms include Principal Component Analysis using eigenfaces, Linear Discriminate Analysis, Elastic Bunch Graph Matching using the Fisher face algorithm, the Hidden Markov model, the Multilinear Subspace Learning using tensor representation, and the neuronal motivated dynamic link matching. System use YCbCr color space model to define the location of face in the image. Values of Cb and Cr component support system in defining the skin part in the input frames as we shown in

algorithm 1. We use the skin part in defining face location and drawing rectangle around it as shown in figure 4.

B. Tracking facial features

After detecting face location system use manual way to track facial features. System able to detect eyes, nose and mouth By using vertical projection in the upper and lower part of detected face as shown in algorithm 1 and figure4.

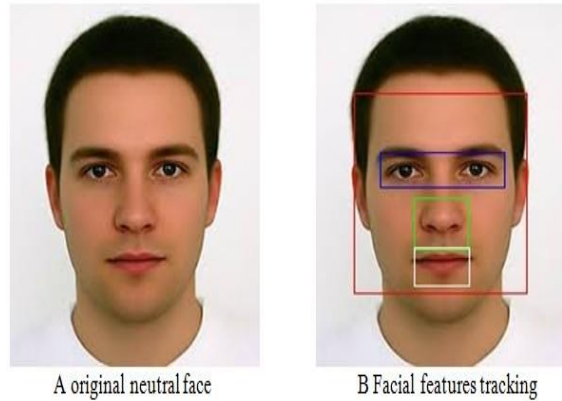


Fig. 4: Extracting facial features

VI. CLASSIFICATION

System stores dictionary for extracted features of Arabic sign language in database. That dictionary supports it in classification. Also in database contains extracted features of face emotions. We use in classification two classifier minimum distance classifier and sum of absolute difference classifier.

A. Minimum Distance Classifier

The minimum distance classifier (MDC) is an example of known used conventional classifier [12], [13]. The single nearest neighbor technique completely bypass the problem of probability distance and simply classifies any unknown sample as belonging to the same class of the most similar or nearest Feature vector in the training set of data. Nearest will be taken to the smallest Euclidean distance in dimensional feature space and the classifier compares the extracted new feature vector $x(i)$ with all the class known feature vectors $y(i)$ and choose the class that minimizes the distance classifier using equation 9

$$Distance = \sum_{i=1}^N |(y(i) - x(i))| \quad (9)$$

Where N is the feature vector length

B. Sum of absolute difference classifier

Sum of absolute difference classifier is considered as a single nearest. It depends on absolute distance between the new feature vector $x(i)$ with all the class known feature vectors $y(i)$ using equation number 10.

$$Distance = \sum_{i=1}^N (y(i) - x(i))^2 \quad (10)$$

Where N is the feature vector length

VII. EXPERIMENTAL RESULTS

In the first part in the model, we use Arabic dictionary for all alpha characters as shown in figure 5. To evaluate the performance of the first part, several videos containing sequences of letters such as "Noon, Ayn, Miem" to generate "Nam" word and "la" have been classified. The system detected the "Noon, Ayn, Miem, la" and generate "Nam" word and "la" word. It has accuracy 99% as we shown in table I. It has the best accuracy comparing with other systems accuracy as we shown in figure 6.

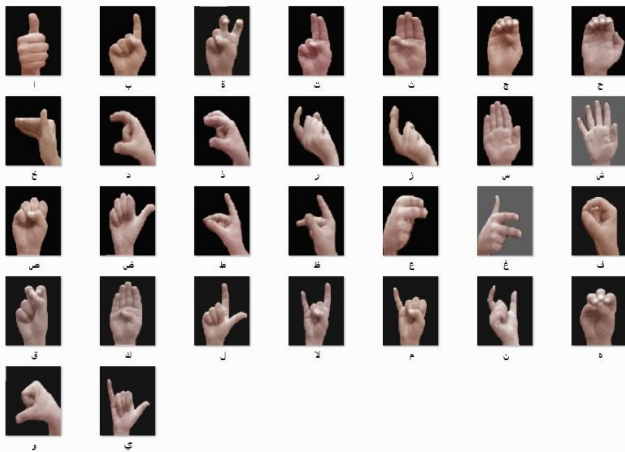


Fig. 5: Arabic signs alpha dictionary

TABLE I: ARABIC SIGN LANGUAGE RECOGNITION

Sign language	Classifier	Recognition rate
Arabic Sign Language	MDC	91.3%
Arabic Sign Language	multilayer perceptron	83.7%
Video-based [3]	hidden Markov	93.8%
Our paper	MDC	99%
Our paper	ADC	99%

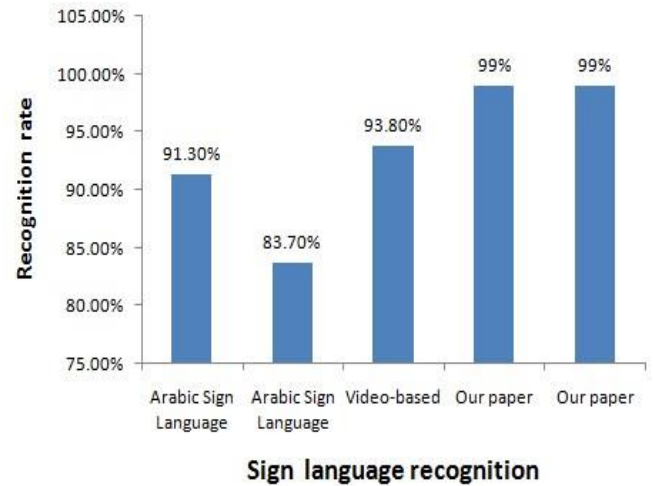


Fig. 6: Arabic sign language recognition rate

In the second part of the system, we are trying to detect face in the image. For face detection we use OpenCv library to support us in defining face location. OpenCV is released under a BSD license; OpenCV is free for both academic and commercial using. It has C++, C, Python and Java interfaces and supports Windows, Linux, Mac OS, iOS and Android. OpenCV was designed for computational efficiency and with a strong support for real-time applications. OpenCv. Implementation in C++/C library can take advantage of multi-core processing. Enabled with OpenCL, it takes advantage of the hardware acceleration of the underlying heterogeneous compute platform. Adopted all around the world or a video frame from a video source. One of its common ways to do this is comparing selected facial features from the image and a facial database. Some of popular facial recognition algorithms identify facial features by extracting landmarks, or features, from an image of the face. For example, an algorithm may analyze the relative position, size, and/or shape of the eyes, nose, cheekbones, and jaw these features are then used to search for other images with matching features. There are other algorithms that depend on normalizing a gallery of face images and compress the face data, save only the data in the image that is important for face recognition. A probe image is then compared with the face data. One of the most successful systems is depended on template matching techniques applied to a set of salient facial features providing a sort of compressed face representation. In that system we use OpenCV library which contains haar cascade frontal face objects that depending on popular algorithm for defining face location. OpenCV support us in defining eyes, mouse and nose location. Depending on defined locations system able to generate feature vector of width and height of each feature location. System deals with seven facial expressions neutral, smile, sad, angry, afraid, disgusted and surprised. Wetake105straightsamplesfromTheKarolinska Directed Emotional Faces (KDEF) for training and 30 straight samples

for testing. KDEF is a set of totally 4900 pictures of human facial expressions of emotion. The material was developed in 1998 by Daniel Lundqvist, Anders Flykt and Professor Arn at Karolinska Institutet, Department of Clinical Neuroscience, Section of Psychology, Stockholm, Sweden. We took 15 training samples for each emotion as we shown in figure 7 the training samples of smile facial expression.

In case of testing, we use 30 samples from KDEF database for testing. Result of testing by using minimum distance classifier was 90% and also 90% by using absolute difference classifier as we shown in table II and table III. System has the best accuracy comparing with other systems as we shown in figure 8 and table IV.

TABLE II: ERROR MATRIX

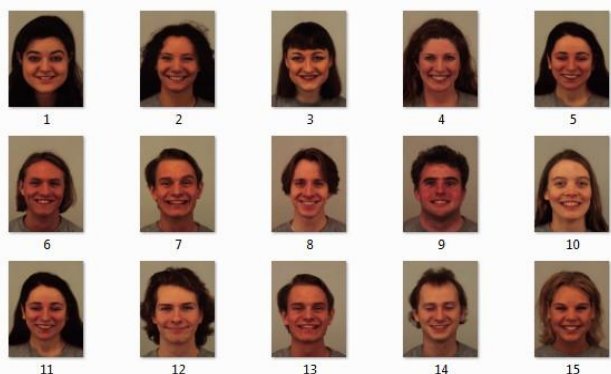


Fig. 7: Training samples of smile face emotion

	Neutra	Smile	Sad	Angry	Afraid	Disgusted	Surprise
Neutral	4	0	0	0	0	0	1

Training Tracker	Testing tracker	Number	Recognition rate
KLT tracker[14]	KLT tracker	4	76%
Manual tracker[14]	KLT tracker	4	63%
Manual tracker[14]	Bayes tracker	4	66%
Bayes tracker[14]	Bayes tracker	4	82%
Manual tracker[14]	Manual tracker	4	84%
Our paper tracker	Our paper tracker	7	90%

Smile	0	4	0	0	0	0	0
Sad	0	0	5	0	0	0	0
Angry	0	0	0	4	1	0	0
Afraid	0	0	0	1	1	0	0
Disgust	0	0	0	1	1	3	1
Surpris	1	0	0	0	0	1	6

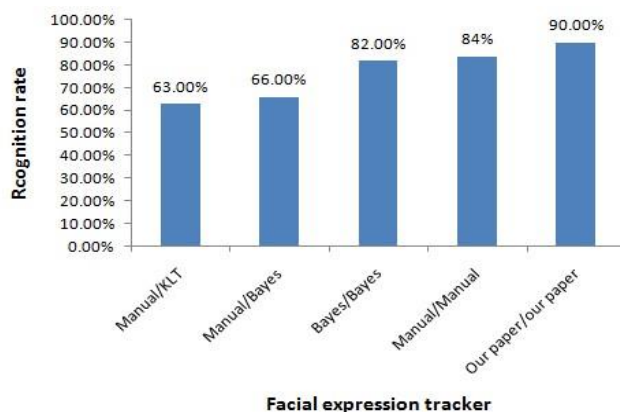


Fig.8: Facial expressions tracker.

TABLE III: TESTING INPUT FACIAL EXPRESSION

TABLE IV: RECOGNITION RATE THAT RESULT FROM USING DIFFERENT FACIAL EXPRESSION TRACKER IN TRAINING AND TESTING.

TABLE V: FINAL DECISION DEPENDING ON THE FINAL RESULT FROM FIRST SUBSYSTEM OF SIGNS AND SECOND SUBSYSTEM OF FACIAL EXPRESSIONS

Emotion	Signs	Final decision	Signs	Final decision
Neutral	Nam	Nam	La	La
Smile	Nam	Nam	La	Nam
Sad	Nam	La	La	La
Angry	Nam	La	La	La

Emotion	No of tested samples	Results	Error
Neutral	5	4	one
Smile	4	4	zero
Sad	5	5	zero
Angry	4	5	one
Afraid	2	1	one
Disgusted	4	3	one
Surprised	6	8	one

VIII. CONCLUSION AND FUTURE WORKS

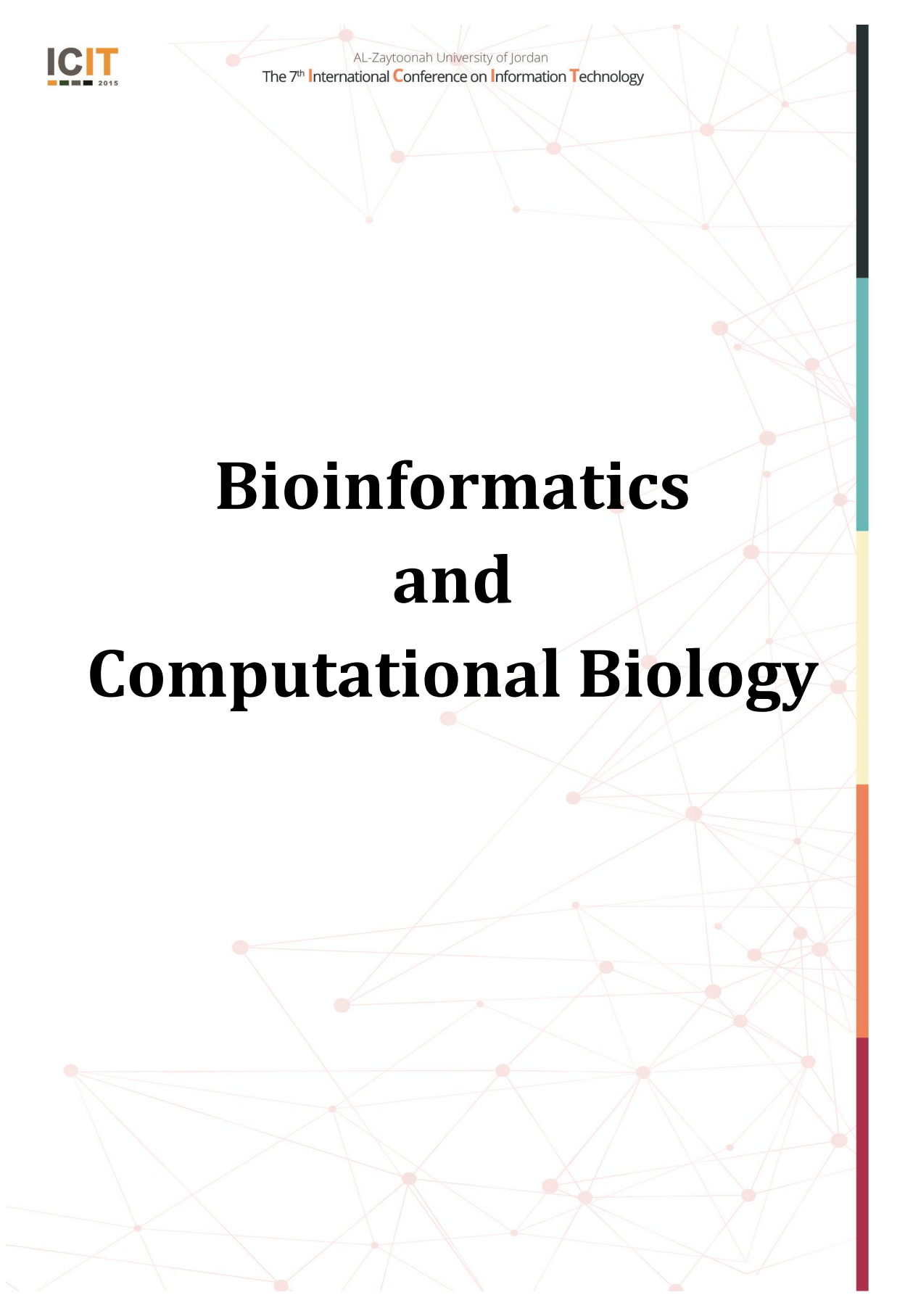
In this paper, a system for the purpose of the recognition and translation of the alphabets in the Arabic sign language was implemented. The system will support deaf people in

interacting with environment as normal people. The system will enable deaf be to transfer their expressions and emotions to others. The system takes facial expressions in its considerations in translation. The system consists of two parts, first part for manual signs and consists of three phases preprocessing phase, skin detection phase and feature extraction phase. Second part in the system that deals with facial expressions consist of two phases detects face and tracking facial expression. System has an accuracy of 90% using minimum distance classifier (MDC) and absolute difference classifier in facial expressions extraction and 99% in case of signs extraction. In the future we will add additional improvements to system to be used for mobile applications to provide easy communication way among deaf/hearing-impaired people. We also could be developed to be provided as a web service used in the field of conferences and meetings attended by deaf people. That system can be used in intelligent class rooms and intelligent environments for real-time translation for sign language. We can support system with other facial expressions like afraid, disgusted and surprised. Common grammatical expressions like Yes/no question (YN), Wh question (WH), Topic (TP), and Negation (NEG) can be developed and add to system to save time and add more supporting to deaf people in their communication. We will increase the size of the database for training and testing. We will use different direction of faces to represent facial expression in 3D. We will use others classifiers in testing. We will add more features to the system to deal with different words that have different meaning if facial expressions changed.

REFERCES

- [1] Nashwa El-Bendary¹, Hossam M. Zawbaa², Mahmoud S. Daoud², Aboul Ella Hassanien², and Kazumi Nakamatsu³, "Arabic Sign Language Al- phabets Translator", International Journal of Computer Information Sys- tems and Industrial Management Applications. Vol. 3, No. 2, PP. 498-506, 2011.
- [2] O. Al-Jarrah and A. Halawani. "Recognition of Gestures in Arabic Sign Language Using Neuro-Fuzzy Systems", Artificial Intelligence, Vol. 133, No. 1-2, PP. 117-138, 2001.
- [3] M. AL-Rousan, K. Assaleh, and A. Tala. "Video-based Signer-independent Arabic Sign Language Recognition Using Hidden Markov Models", Applied Soft Computing, Vol. 9, No. 3, PP. 990-999, 2009
- [4] V. I. Pavlovic, R. Sharma, and T. S. Huang. "Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review", IEEE Trans. Pattern Anal. Machine Intell. , Vol. 19, No. 7, PP. 677-695, 1997
- [5] M. Huenerfauth. "Generating American Sign Language Classifier Predicates For English-To ASL Machine Translation", Ph.D dissertation, University of Pennsylvania, Department of Computer and Information Science, Philadelphia, PA, USA, 2006.
- [6] S. M. Halawani. "Arabic Sign Language Translation System on Mobile Devices", International Journal of Computer Science and Network Security (IJCSNS), Vol. 8, No. 1, PP. 251-256, 2008
- [7] J.A. Bangham, S.J. Cox, M. Lincoln, ITutt, and M. Wells. "Signing for the Deaf Using Virtual Humans", IEE Seminar on Speech and Language Processing for Disabled and Elderly People, No. 2000/025, PP. 4/1-4/5, 2000.
- [8] S. Phitakwinai, S. Auephanwiriyaikul, and N. Theera-Umpon. "Thai Sign Language Translation Using Fuzzy C-Means and Scale Invariant Feature Transform", The Proceedings of International Conference of Computational Science and Its Applications, PP. 1107-1119, Thai, June 2008.
- [9] A. Kanaujia and D. N. Metaxas. "Large Scale Learning of Active Shape Models". The Proceedings of IEEE International Conference on Image Processing, PP. 265-268, San Antonio, 16-19 Sept. 2007.
- [10] Y. Tong, Y. Wang, Z. Zhu, and Q. Ji. Robust "facial feature tracking under varying face pose and facial expression" .Pattern Recognition, Vol. 40, No. 40, PP. 3195-3208, 2007.
- [11] Y. Tian, T. Kanade, and J. Cohn. "Recognizing Action Units for Facial Expression Analysis". IEEE transactions on Pattern Analysis and Machine Intelligent, Vol. 23, No. 2, PP 97-115, 2001.
- [12] M. S. Packianather and P. R. Drake. "Comparison of Neural and Minimum Distance Classifiers in Wood Veneer Defect Identification", The Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture, Sage Publications, Vol. 219, No. 11, PP. 831, 2005.
- [13] R. Boveiri. "Persian Printed Numerals Classification Using Extended Moment Invariants", The proceedings of WASET Int. Conf. on Image and Vision Computing de Janeiro, PP. 167-74, World Academy of Science, Engineering and Technology, 2010.
- [14] R. Boveiri. Tan Dat Nguyen and Surendra Ranganath, "Tracking facial features under occlusions and recognizing facial expressions in sign language", The proceeding of IEEE International Conference, PP. 1 7, Amsterdam, 17-19 Sept. 2008.
- [15] Mr. Manoj K. Vairalkar and Prof. S.U. Nimbhorka. "Edge Detection of Images Using Sobel Operator", International Journal of Emerging Technology and Advanced Engineering, Vol. 2, No. 1, PP. 2250-2459, 2012

Bioinformatics and Computational Biology



Comparison between X-Ray Radiography Image Fusion Algorithms Used in Medical Applications

Mohamed S. El_Tokhy

Electrical Engineering Department, College of Engineering
Aljouf University
Aljouf, KSA
engtokhy@gmail.com

Ibraheem M. Fayed

Electrical Engineering Department, College of Engineering
Aljouf University
Aljouf, KSA
imfayed@ju.edu.sa

Abstract— Medical diagnoses methods are very important to determine the reasons of diseases. In this paper different algorithms are introduced to make image fusion using x-ray radiography chest image. This fusion is done for front view image and side view image. These algorithms are multiplication factors (MF), multi-resolution singular value decomposition (MSVD), dual tree complex wavelet transform, and discrete stationary wavelet transform (SWT). These algorithms are considered using six different methods. These methods being averaging, max coefficient, block bias based on largest magnitude, energy, block bias based on largest contrast and bias methods. A comparison is accomplished using peak signal to noise ratio (PSNR), mean square error (MSE), entropy, and measure of structural similarity (SSIM). The obtained results confirms that applying the fusion algorithms introduces better results for the four algorithms and enhances the performance characteristics of these algorithms more than using front view image only or side view image only. Furthermore, we concluded that the conventional MF algorithm is superior other three algorithms for most of statistical characteristics.

Keywords- image processing, fusion algorithm, and radiography.

I INTRODUCTION

+Biomedical image processing is a rapidly growing area of research from last two decades. Availability of numerous kinds of biomedical sensors has increased the interest of researchers and scientists in this field. X-ray, ultrasound, magnetic resonance imaging (MRI) and computed tomography (CT) are a few examples of biomedical sensors. These sensors are used for extracting clinical information, which are generally complementary in nature. For example, X-ray is widely used in detecting fractures and abnormalities in bone position, CT is used in tumor and anatomical detection and MRI is used to obtain information among tissues. Similarly, other functional imaging techniques like functional magnetic resonance imaging (MRI), positron emission tomography (PET), and single positron emission computed tomography (SPECT) provide functional and metabolic information. Hence, none of these modalities is able to carry all relevant information in a single image. Therefore, multimodal fusion is required to obtain all possible relevant information in a single composite image [1]. The fusion of data for medical imaging has become a central issue in such biomedical applications as image-guided surgery and radiotherapy [2]. Image fusion is a process to combine information from multiple images of the same scene [3], [4]. The result of image fusion will be a new

image which is more suitable for human and machine perception or further tasks of image processing such as image segmentation, feature extraction and object recognition [3]. There are two basic requirements for image fusion [1].

- Fused image should possess all possible relevant information contained in the source images;
- Fusion process should not introduce any artifact, noise or unexpected feature in the fused image.

Image Fusion is one of the important and preprocessing steps in digital image reconstruction [5]. The objective of image fusion is to better the quality of fused images, extract all the useful information from the source images and do not introduce artifacts or inconsistencies which will distract human observers. Many algorithms have been developed for fusion of medical images as reported in the literature [5]. Despite the significant research conducted on this topic, the development of efficient medical image fusion method is still a big challenge for the researchers [5]. This comparison is accomplished using peak signal to noise ratio (PSNR), mean square error (MSE), entropy, and measure of structural similarity (SSIM).

II IMAGE FUSION ALGORITHMS

The fusion process should preserve all relevant information in the fused image, should reduce noise and should suppress any artifacts in the fused image [6], [7].

Image fusion is the process of integrating all relevant and complementary information from different source images into a single composite image without introducing any artifact or noise [1]. Image fusion can be performed at three levels; pixel level, feature level and decision level [1], [3]. Pixel level fusion deals with information associated with each pixel and fused image can be obtained from the corresponding pixel values of source images [1]. In feature level fusion, source images are segmented into regions and features like pixel intensities, edges or texture, are used for fusion. Decision level fusion is a high level fusion which is based on statistics, voting, fuzzy logic, prediction and heuristics [1].

In the field of image fusion, pixel-level fusion becomes the primary method since it can preserve original information of source images as much as possible, and the algorithms are computationally efficient and easy to implement, the most image fusion applications employ pixel level based method [3]. There are three commonly used methods of pixel-level image fusion, including simple image fusion [3] (such as linear weighted average, HPF (high-pass-filter), HIS (intensity hue-saturation), PCA (principal component analysis)), pyramid-based decomposition image fusion (such as Laplace pyramid decomposition, ratio pyramid) and wavelet transform image fusion [3]. Recently, wavelet transform becomes an important aspect of image fusion research with the merits of multi-scale and multi-resolution [3].

Therefore, the radiography chest image fusion from different views is taken as an example to introduce better diagnoses. Therefore, three algorithms are evaluated and used for this purpose.

A. X-ray radiography chest image fusion based on multiplication factor

The fusion of radiography chest images can be realized in successive steps as depicted in Fig. 1. This algorithm is based on a MATLAB routine implementing the image fusion algorithm in [8]. In this algorithm, we combine two x-ray radiography chest images. One of the advantages of this algorithm, It supports both gray and color images. The basic idea of this algorithm depends on factor that is lies between zero and one. This factor can be varied to vary the proportion of mixing of each image. Therefore, there are three different cases dependent on this factor (F):

$$F = \begin{cases} = 0.5 & \text{Equal Mixing} \\ < 0.5 & \text{Side View Image Contribution} \\ > 0.5 & \text{Front View Image Contribution} \end{cases} \quad (1)$$

The side view image is multiplied by this factor. However, the front end image was multiplied by (1-F). Both images are added to obtain the fused image. The contribution

of both images based on the value of this factor. If the value of the factor is equal to 0.5, contribution of side view and front view images are identical. Contribution of side view image is larger as this value smaller than 0.5. However, contribution of front view image is larger as this value larger than 0.5.

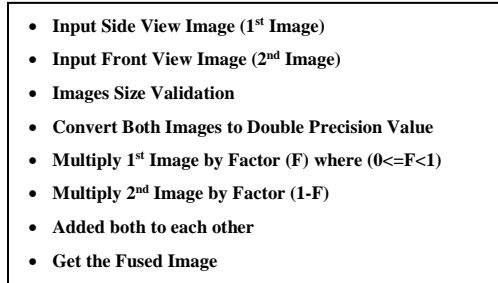


Figure 1. X-ray radiography chest image fusion algorithm using multiplication factor

B. X-ray radiography chest image fusion based on multi-resolution singular valuedecomposition

Multi-resolution singular value decomposition (MSVD) is very similar to wavelets transform, where signal is filtered separately by low pass and high pass finite impulse response (FIR) filters and the output of each filter is decimated by a factor of two to achieve first level of decomposition [9]. The decimated low pass filtered output is filtered separately by low pass and high pass filter followed by decimation by a factor of two provides second level of decomposition. The successive levels of decomposition can be achieved by repeating this procedure. The idea behind the MSVD is to replace the FIR filters with singular value decomposition (SVD) [9]. An algorithm for image fusion based on MSVD is studied as depicted in Fig. 2. This algorithm is based on a MATLAB routine implementing the image fusion algorithm in [9].

The images to be fused I1 and I2 are decomposed into L(l=1,2,...,L) levels using MSVD. At each

decomposition level (l =1,2,..., L) , the fusion rule will select the larger absolute value of the two MSVD detailed coefficients, since the detailed coefficients correspond to sharper brightness changes in the images [9]. These coefficients are fluctuating around zero. At the coarsest level (l = L), the fusion rule take average of the MSVD approximation coefficients since the approximation coefficients at coarser level are the smoothed and subsampled version of the original image. Similarly, at each decomposition level (l =1,2,..., L), the fusion rule take the average of the two MSVD eigen matrices [9].

- **Input Side View Image (1st Image)**
- **Input Front View Image (2nd Image)**
- **Images Size Validation**
- **Apply Multi-Resolution Singular Value Decomposition**
 - Images are decomposed into different levels
 - At each level, choose the largest absolute value of the two MSVD detailed coefficients
 - At coarsest level, take average of the MSVD approximation coefficients
 - Select average of the two MSVD eigen matrices at each decomposition level
- **Take Inverse of Multi-Resolution Singular Value Decomposition**
 - Extract the spatial domain image from the MSVD coefficients

Figure 2. X-ray radiography chest image fusion algorithm using multi-resolution singular value decomposition

C. *X-ray radiography chest image fusion based on dual tree complex wavelet transform*

Another algorithm for image fusion based on dual-tree complex wavelet transform is studied as depicted in Fig. 3. This algorithm is based on a MATLAB routine implementing the image fusion algorithm in [10]. There are two representations of the 2D dual-tree wavelet transform; the real 2D dual-tree DWT and complex 2D dual-tree DWT. The dual-tree complex wavelet transform (DTCWT) is a relatively recent enhancement to the DWT, with important additional properties: nearly shift-invariant and directionally selective (useful in two and higher dimensions) [11], [12]. In the dual-tree implementation of decomposition and reconstruction, two parallel DWTs with different low-pass and high-pass filters in each scale are used as can be seen in Fig. 4 [11]. The two DWTs use two different sets of filters, with each satisfying the perfect reconstruction condition.

One of the advantages of the dual-tree complex wavelet transform is that it can be used to implement 2D wavelet

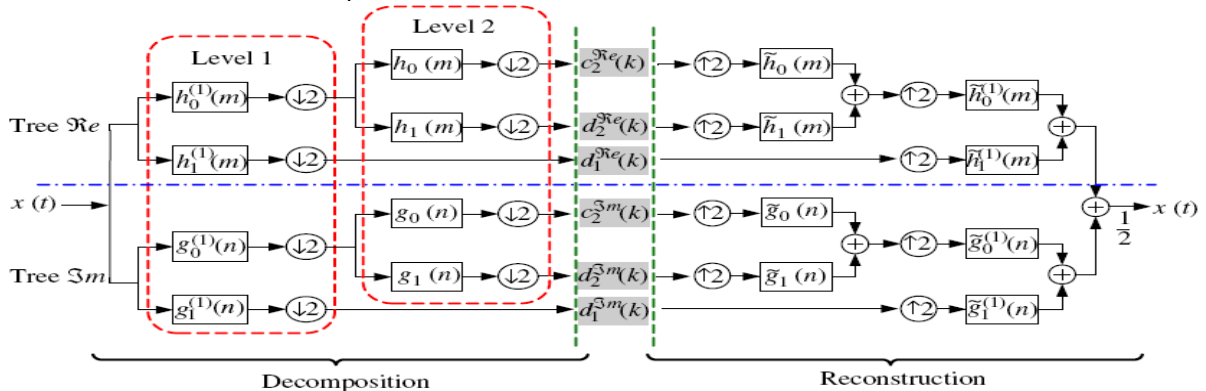


Figure 4. Two-stage DTCWT decomposition and reconstruction.

D. *X-ray radiography chest image fusion based on discrete stationary wavelet transform*

Recently, the DWT has become a powerful tool for multiscale image fusion [13]. Stationary wavelet transform (SWT) is similar to discrete wavelet transform (DWT) but the only process of down-sampling is suppressed that means the SWT is translation-invariant [4]. It is redundant, shift

transforms that are more selective with respect to orientation than is the separable 2D discrete wavelet transform (DWT) [10], [11]. The complex 2D dual-tree DWT gives rise to wavelets in six distinct directions. In each direction, one of the two wavelets can be interpreted as the real part of a complex-valued 2D wavelet, while the other wavelet can be interpreted as the imaginary part of a complex-valued 2D wavelet. The complex 2D dual-tree is implemented as four critically-sampled separable 2D DWTs operating in parallel [10]. However, different filter sets are used along the rows and columns [10], [11]. Furthermore, the sum and difference of sub-band images is performed to obtain the oriented wavelets [10].

- **Input Side View Image (1st Image)**
- **Input Front View Image (2nd Image)**
- **Use Farras filter and Kingsbury Q-Filters to Analyze Filters for Tree i**
- **Select the Number Of Decomposition Levels**
- **Make Image Decomposition Using Dual Tree Complex 2D Discrete Wavelet Transform**
- **Determine Real and Imaginary Parts of These Deduced Coefficients**
- **Apply Inverse Dual Tree Complex 2D Discrete Wavelet Transform**
- **Extract the Fused Image**

Figure 3. X-ray radiography chest image fusion algorithm using dual tree complex discrete wavelet transforms

invariant, and gives a more dense approximation to the continuous wavelet transform than discrete wavelet transforms [14]. The discrete wavelet transform (DWT) is a common tool for image fusion, but the result could contain the artifacts near the edges. This impairment is addressed by the models based on stationary wavelet transform (SWT), curvelet transform and non-sampled contourlet (NSCT). Recent studies show that both SWT and NSCT turn out to be

the more suitable fusion approaches because of their shift-invariant [15]. The way to restore the translation invariance is to average some slightly different DWT, called decimated DWT, to define the stationary wavelet transform (SWT). Let us recall that the DWT basic computational step is a convolution followed by decimation. The decimation retains even indexed elements. But the decimation could be carried out by choosing odd indexed elements instead of even indexed elements [6].

The SWT algorithm is very simple and is close to the DWT one [4], [6], [13]. More precisely, for level 1, all the decimated DWT for a given signal can be obtained by convolving the signal with the appropriate filters as in the DWT case but without down sampling. Then, the approximation and detail coefficients at level 1 are both of size N, which is the signal length. The general step j convolves the approximation coefficients at level j-1, with up sampled versions of the appropriate original filters, to produce the approximation and detail coefficients at level j [6].

Figure 5 and Figure 6 Show the fusion process for the front view image and the side view image to get better image with more details.

- **Load Side View Image (1st Image)**
- **Load Front View Image (2nd Image)**
- **Perform Image decomposition using discrete stationary wavelet transforms into different levels**
- **Display the approximation and detail coefficients at the different levels**
- **Reconstruct the fused image from these coefficients by Applying the Inverse of discrete stationary wavelet transforms**

Figure 5. X-ray radiography chest image fusion algorithm using discrete stationary wavelet transforms

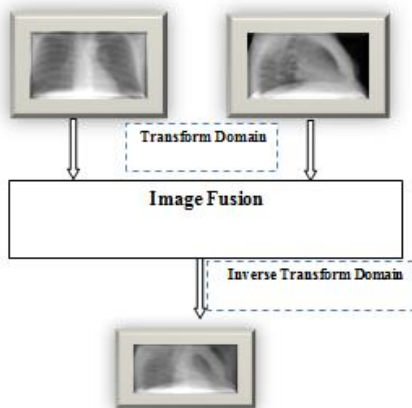


Figure 6. Image fusion algorithm based on DST and FFT

III RESULTS AND DISCUSSIOA

Comparison between these algorithms is of major concern. Therefore, this comparison study is done between the considered algorithms using different statistical evaluation processes. These, statistics are root mean square error (RMSE), mean absolute error (MAE), signal to noise ratio (SNR), and peak signal to noise ratio (PSNR). The statistical measurements between fused image and both front image and side image for multiplication, MSVD, DTCWT and SWT algorithms are depicted in Tables 1-4. We noticed that MSE between fused image and side view image is lower than that between fused image and front one for most algorithms. Also, SNR and PSNR between fused image and side view image are larger than that between fused image and front one for most algorithms.

	MSE	PSNR	Entropy	SNR	RMSE	MAE
MF for Front Image	39.7887	32.1332	7.477	0.00670	6.3078	0.1952572
MF Side Image	22.8838	34.5355	7.477	1.02669	4.7837	0.1952572

	MSE	PSNR	Entropy	SNR	RMSE	MAE
MSVD for Front Image	0.029093	63.4929	6.8489	0.15495	0.17057	0.13495
MSVD Side Image	0.028496	63.5829	6.8489	0.69227	0.16881	0.13426

	MSE	PSNR	Entropy	SNR	RMSE	MAE
DTCWT for Front Image	0.082808	58.9501	6.6614	2.54017	0.28776	0.20825
DTCWT for Side Image	0.17579	55.6809	6.6614	3.07749	0.41927	0.31962

	MSE	PSNR	Entropy	SNR	RMSE	MAE
SWT for Front Image	0.029068	63.4966	6.8945	0.68034	0.17049	0.13479
SWT for Side Image	0.028382	63.6003	6.8945	0.14631	0.16847	0.13388

Comparison between multiplication factor (MF), MSVD, DTCWT, and SWT(L1) algorithms based on the mentioned statistics before and after applying the fast lifting transform are depicted in Tables 1-2, respectively. From these tables, the discrete stationary wavelet transform (SWT) and MF achieve the best performance for most characteristics before and after applying the fast lifting transform.

Statistics	MF	MSVD	DTCWT	SWT(L1)
RMSE	0.2473732	0.2216372	0.2858792	0.1704952
FE (%)	46.9770992	53.3529622	54.2895132	30.0690352
MAE	0.1952572	0.1852432	0.2225302	0.1347882
CORR	0.8892772	0.8900532	0.8124702	0.9512592
SNR	6.5622762	5.4568292	5.3056812	10.4376102
PSNR	54.2312812	54.7083792	53.6029832	55.8476892
MI	1.0807172	1.0016432	1.0977672	1.0245702
QI	0.1177012	0.1208092	0.2189232	0.5043482
SSIM	0.9927212	0.9935042	0.9895282	0.9968282

Statistics	MF	MSVD	DTCWT	SWT(L1)
RMSE	0.2473732	0.2471542	0.2858992	0.1684962
FE (%)	46.9770992	46.9200632	54.2754102	31.5915932
MAE	0.1952572	0.1952642	0.2227982	0.1339862
CORR	0.8892772	0.8893332	0.8125132	0.9492692
SNR	6.5622762	6.5728282	5.3079382	10.0085692
PSNR	54.2312812	54.2351182	53.6026732	55.8989042
MI	1.0807172	1.0820882	1.0951482	1.1495912
QI	0.1177012	0.1214162	0.2188072	0.4466332
SSIM	0.9927212	0.9927562	0.9895242	0.9967232

IV CONCLUSION

Different algorithms are evaluated to make image fusion using x-ray radiography chest image. This fusion is done for front view image and side view image. These algorithms are

named multiplication factor (MF), multi-resolution singular value decomposition (MSVD), dual tree complex wavelet transform, and discrete stationary wavelet transform (SWT). These algorithms are evaluated using statistical measurements. The obtained results confirm that MSE between fused image and side view image is lower than that between fused image and front one for most algorithms. Also, SNR and PSNR between fused image and side view image are larger than that between fused image and front one for most algorithms. Therefore, using the fused image and side view image will have more details than front image with fused image. Furthermore, using fused image and side view difference processing with SWT and MF makes diagnosis of respiratory diseases more accurate.

REFERENCES

- Rajiv Singh, Ashish Khare, "Fusion of multimodal medical images using Daubechies complex wavelet transform-A multiresolution approach", *Information Fusion*, Vol. 19, pp. 49-60, 2012 [http://dx.doi.org/10.1016/j.inffus.2012.09.005].
- Zhiping Xu, "Medical image fusion using multi-level local extrema", *Information Fusion*, Vol. 19, pp.38-48, 2013 [http://dx.doi.org/10.1016/j.inffus.2013.01.001]
- S. V. More, and S. D. Apte, "Pixel-Level Image Fusion Using Wavelet Transform", *International Journal of Engineering Research & Technology (IJERT)*, Vol. 1, Vol. 5, 2012.
- Pusit Borwonwatanadelok, Wirat Rattanapitak and Somkait Udomhunsakul, "Multi-Focus Image Fusion based on Stationary Wavelet Transform and extended Spatial Frequency Measurement", *International Conference on Electronic Computer Technology*, 2009.
- Navneet Kaur, Jaskiran Kaur, "A Novel Method For Pixel Level Image fusion Based on Curvelet Transform", *International Journal of Research in Engineering and Technology (IJRET)*, Vol. 1, No. 1, 2013.
- Kanagaraj Kannan, Subramonian Arumuga Perumal, Kandasamy Arulmozhi, "Optimal decomposition level of discrete, stationary and dual tree complex wavelet transform for pixel based fusion of multi-focused images", *Serbian Journal of Electrical Engineering*, Vol. 7, No. 1, May 2010, 81-93
- Y. Chai, H.F. Li, M.Y. Guo, "Multifocus image fusion scheme based on features of multiscale products and PCNN in lifting stationary wavelet domain", *Optics Communications*, Vol. 284, pp. 1146-1158, 2011.
- Athi, Matlab Central, "mixing or combining two images (image fusion)", Mathwork, 2009.
- V.P.S. Naidu, "Image Fusion Technique using Multi-resolution Singular Value Decomposition", *Defence Science Journal*, Vol. 61, No. 5, , pp. 479-484, 2011.
- V.P.S. Naidu, "2D dual tree complex wavelet transform", http://taco.poly.edu/WaveletSoftware/, 2013.
- Yanxue Wang, Zhengjia He, Yanyang Zi, "Enhancement of signal denoising and multiple fault signatures detecting in rotating machinery using dual-tree complex wavelet transform", *Mechanical Systems and Signal Processing*, Vol. 24, pp. 119-137, 2010.
- Ivan W. Selesnick, Richard G. Baraniuk, and Nick G. Kingsbury, "The Dual-Tree Complex Wavelet Transform", *IEEE Signal Processing Magazine*, pp. 123-151, 2005.
- Hailiang Shi, Min Fang, "Multi-focus Color Image Fusion Based on SWT and IHS", *Fourth International Conference on Fuzzy Systems and Knowledge Discovery*, 2007.
- Cungui Cheng, Jia Liu, Wenqing Cao, Renwei Zheng, Hong Wang, Changjiang Zhang, "Classification of two species of Bidens based on discrete stationary wavelet transform extraction of FTIR spectra

combined with probability neural network”, *Vibrational Spectroscopy*, Vol. 54, No. 1, pp. 50-55, 2010.

- [15] Zhou Zeming, Jiang Lin, Wang Jin, Zhang Peng, Yang Pinglv, “Image Fusion by Combining SWT and Variational Model”, 4th International Congress on Image and Signal Processing, Vol. 4, 2011.

Detecting patients with Parkinson's disease using PLP and VQ

Achraf Benba

Laboratoire de Recherche en Génie Electrique, Ecole Normale Supérieure de l'Enseignement Technique,
Mohammed V University
Rabat, Morocco
achraf.benba@um5s.net.ma

Abdelilah Jilbab

Laboratoire de Recherche en Génie Electrique, Ecole Normale Supérieure de l'Enseignement Technique,
Mohammed V University
Rabat, Morocco

Ahmed Hammouch

Laboratoire de Recherche en Génie Electrique, Ecole Normale Supérieure de l'Enseignement Technique,
Mohammed V University
Rabat, Morocco

Abstract— Parkinson's disease (PD) is a neurological disorder of unknown etiology. PD causes several symptoms during its course, and this includes voice disorders of 90% of patients. In order to improve the evaluation of these disorders, we have used 34 voice samples of sustained vowel /a/, from a set of 34 people including 17 patients with PD. We subsequently extracted from each person, from 1 to 20 coefficients of the Perceptual linear prediction (PLP). The frames of the PLP were compressed using vector quantization, with six codebook sizes namely; 1, 2, 4, 8, 16 and 32. We used the technique of Leave One Person Out (LOPO) and the Support Vector Machines (SVMs) classifier with two types of kernels; RBF and Linear. The obtained results using the codebook size of 1 were not stable. Therefore, we proceeded to a bench of 100 trials. The best average accuracy obtained was 75.8%, and the maximum classification accuracy obtained was 91.17% using the codebook size of 1.

Keywords— Parkinson's disease, Perceptual Linear Prediction, Vector quantization, Leave One Person Out, Support Vector Machines.

I. INTRODUCTION

THE evaluation of the quality of voice, and the identification of the causes of its degradation based on phonological and acoustic traits have become major concerns of clinicians and voice pathologists. They have become more attentive to any external techniques to their domain, which might provide them additional information for the evaluation of PD. During its course, PD causes different symptoms and affects the system which controls the execution of learned motor plans such as walking, talking or completing other simple tasks [1] [2] [3]. PD is generally seen in people whose age is over 50 years and causes voice weakening in approximately 90% of patients [4]. For these patients, physical visits for diagnosis, monitoring and treatment are too hard [5] [6].

In the case of the evaluation of voice disorders caused by PD, clinicians and the voice pathologists have adopted subjective techniques based on acoustic traits to distinguish different disease levels. In order to develop more objective

evaluations, recent studies use measurements of voice quality in time, frequency and cepstral domains [7] to detect voice disorders in the context of PD. Such as; fundamental frequency of the oscillation of vocal folds, absolute sound pressure level, jitter which represents pitch perturbations, shimmer which represents amplitude perturbations, and harmonicity which represents the degree of acoustic periodicity [1] [8] [9].

In this research we focused on the measurements in cepstral domain by applying PLP cepstral coefficients (CC) which have been usually used in speaker identification applications and were first proposed by H. Hermansky [10]. We have extracted PLP coefficients from the voice signals provided in a dataset and used VQ for data compression. We subsequently used the LOPO validation scheme with SVMs for data classification in order to discriminate PD patients from healthy people.

This paper is organized as follows: the voice samples dataset is described in section II. The PLP processes and VQ are presented successively in section III and IV. The

methodology of this research and the results are presented in section V and conclusion in section VI.

II. DATASET

The data collected in the context of this research belongs to 17 patients with PD (6 women, 11 men) and 17 healthy people (8 women, 9 men). Voice samples were recorded through a standard microphone at a sampling frequency of 44,100 Hz using a 16 bit sound card in a desktop computer. The microphone was placed at a 15 cm distant from people and they were requested to say sustained vowel /a/ at a comfortable level. All the recordings of voice samples were made in mono channel mode and saved in WAV format; acoustic analyses were applied on these voice samples. All the recordings were sent by M. Erdem Isenkul from the Department of Computer Engineering at Istanbul University, Istanbul, Turkey.

III. PLP PROCESSES

Our first purpose was to transform the voice signal to some type of parametric representation for more analysis and processing [13]. The voice signal is a slow time varying signal which is called quasi-stationary [13]. When it is observed over a short period of time, it appears fairly stable [13]. However, over a long period of time, the voice signal changes its shape.

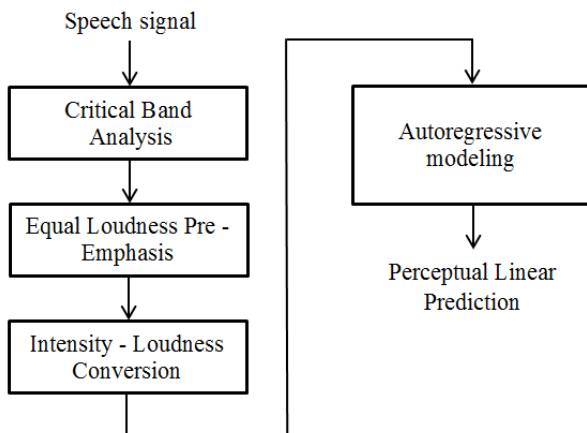


Fig 1: Block diagram of PLP coefficients process

Therefore, it should be characterized by doing short-time spectral analysis [13]. The process of calculating the PLP is shown in Figure 1 and described in the following paragraphs.

A. Spectral Analysis

Since the voice signal is a real signal, it is finite in time; thus, a processing is only possible on finite number of samples [14]. Therefore, the first phase of PLP process is to weight the voice segment by Hamming window [10]. The aim is to reduce signal discontinuities, and make the ends smooth enough to connect with the beginnings [14]. This was

achieved by using Hamming window to taper the signal to zero in the beginning and in the end of each frame, by applying the following equation to the samples [10]:

$$W(n) = \left\{ 0,54 - 0,46 \cdot \cos\left(\frac{2\pi n}{N-1}\right) \right\} \quad (1)$$

where N is the length of the Hamming window, with a length about 20 ms.

The next processing phase consists on converting each frame of N samples from time domain into frequency domain by applying the Fast Fourier Transform (FFT) [13]. We used the FFT for the reason that it is a fast algorithm to implement the Discrete Fourier Transform (DFT) [13]. As known, the DFT is defined on the set of N samples S_n as follow [13]:

$$S_n = \sum_{k=0}^{N-1} s_k e^{-2\pi jkn/N}, n = 0,1,2,\dots, N-1 \quad (2)$$

The short-term power spectrum is calculated by adding the square of the real and imaginary components of short-term voice spectrum, as follow [10]:

$$P(\omega) = \text{Re}[S(\omega)]^2 + \text{Im}[S(\omega)]^2 \quad (3)$$

B. Critical Band Analysis

The short-term power spectrum $P(\omega)$ is warped along its frequency axis ω where $\omega=2\pi f$, into Bark frequency Ω by using the following equation [10]:

$$\Omega(\omega) = 6 \ln \left\{ \frac{\omega}{1200\pi} + \sqrt{\left[\left(\frac{\omega}{1200\pi} \right)^2 + 1 \right]} \right\} \quad (4)$$

$$\Omega(f) = 6 \ln \left\{ \frac{f}{600} + \sqrt{\left[\left(\frac{f}{600} \right)^2 + 1 \right]} \right\} \quad (5)$$

$$\Omega(f) = 6 \sinh^{-1} \left(\frac{f}{600} \right) \quad (6)$$

where ω is the angular frequency in [rad/s], and f is the frequency in [Hz]. The aim of the next phase, is to convolve the resulting warped power with the power spectrum of the simulated critical-band masking curve $\Psi(\Omega)$ approximated by H. Hermansky [10] as follow:

$$\Psi(\Omega) = \begin{cases} 0 & \text{for } \Omega < -1,3 \\ 10^{2,5(\Omega+0,5)} & \text{for } -1,3 \leq \Omega \leq -0,5 \\ 1 & \text{for } -0,5 < \Omega < 0,5 \\ 10^{-1,0(\Omega-0,5)} & \text{for } 0,5 \leq \Omega \leq 2,5 \\ 0 & \text{for } \Omega > 2,5 \end{cases} \quad (7)$$

It is a rather curd approximation of the shape of auditory filters.

The samples of the critical-band power spectrum are produced by doing the discrete convolution of $\Psi(\Omega)$ with $P(\omega)$ by applying the following equation [10]:

$$\theta(\Omega_i) = \sum_{\Omega=-1,3}^{2,5} P(\Omega - \Omega_i) \Psi(\Omega) \quad (8)$$

The convolution between the relatively broad critical-band masking curve $\Psi(\Omega)$ and the short-term power spectrum $P(\omega)$, reduces the spectral resolution of $\theta(\Omega)$ in comparison with the original $P(\omega)$ [10].

C. Equal-loudness Pre-emphasis

The next phase in this process is to pre-emphasize the samples $\Theta[\Omega(\omega)]$ using the simulated equal-loudness curve, by applying the following equation [10]:

$$\Xi[\Omega(\omega)] = E(\omega) \times \Theta[\Omega(\omega)] \quad (9)$$

where, $E(\omega)$ is an approximation to the non-equal sensitivity of human ear perception at different frequencies. The practical approximation used in this research was adopted by H. Hermansky [10] and was first proposed by Makhol *et al* [15] and given by the following equation:

$$E(\omega) = \frac{(\omega^2 + 56,8 \times 10^6) \omega^4}{(\omega^2 + 6,3 \times 10^6)^2 \times (\omega^2 + 0,38 \times 10^9)} \quad (10)$$

$$E(f) = \left[\frac{f^2}{f^2 + 1,6 \times 10^5} \right]^2 \times \left[\frac{f^2 + 1,44 \times 10^6}{f^2 + 9,6 \times 10^6} \right] \quad (11)$$

D. Intensity-loudness Power Law

The last operation before the all-pole modeling is the cubic-root amplitude compression. The following equation approximates the power law of human hearing and simulates the non-linear relation between the intensity of sound and its perceived loudness [10]:

$$\Phi(\Omega) = \Xi(\Omega)^{0,33} \quad (12)$$

E. Autoregressive Modeling

In the final phase of the PLP process, $\Phi(\Omega)$ is approximated by the spectrum of an all-pole model using the autocorrelation

technique of all-pole spectral modeling, this technique is called Linear Prediction (LP) [10] [16], in which the signal spectrum is modeled by an all-pole spectrum. In this research we used the Linear Predictive Coefficient (LPC) analysis to calculate the autoregressive model from spectral magnitude samples. The autoregressive coefficients are transformed to CC of the all-pole model; this was achieved by converting the LPC of n coefficients into frames of CC [10].

F. Liftering

The principal advantage of CC is that they are uncorrelated [14]. However, the problem with them is that the higher orders CC are quite small [14]. Therefore, it is essential to re-scale the CC in order to have quite similar magnitudes [14]. This is achieved by liftering the CC according to the following equation [14]:

$$c'_n = \left(1 + \frac{L}{2} \cdot \sin\left(\frac{\pi \cdot n}{L}\right) \right) \cdot c_n \quad (13)$$

where L is the cepstral sine lifter parameter. In this research, we used $L=0.6$.

IV. VECTOR QUANTIZATION

VQ is a compression technique with data-loss [17]. The basic idea of this technique is to take a large number of data vectors and minimize it to a smaller group of data vectors, which represent the centers of gravity of the distribution.

The VQ technique consists of extracting a small number of the most representative data to characterize different people. Here VQ is used to minimize the number of frames of the coefficients of the PLP in order to have only the most significant vectors which represent the center of gravity of the distribution of other frames of the PLP coefficients. In this research, we have made tests using codebook sizes of 1, 2, 4, 8, 16 and 32 [18].

V. METHODOLOGY & RESULTS

The first phase in this research was to build a dataset containing voice samples of patients with PD and healthy people. Ultimately, we were able to collect 17 voice samples from both groups. This gave us 34 voice samples [19]. These recordings were made through a standard microphone at a sampling rate of 44100 Hz. All participants were asked to pronounce the sustained vowel /a/ at a comfortable level.

All the algorithms were executed on a desktop computer with a Core (TM) i3-2120 CPU and a processing speed of 3.30 GHz. We subsequently extracted from each voice sample, CC of the PLP. The number of PLP coefficients extracted ranged from 1 to 20. We proceeded in this way to get the optimal coefficient number needed for the best classification accuracy.

The PLP coefficients extracted from each sample contains a large number of frames which require extensive processing time for classification and prevents making the correct diagnostic decision. To overcome this problem, and reduce the

processing time, we used a technique of compression with data-loss known as VQ. The detailed description of this technique has been made in section IV. As we know, VQ compresses the frames according to the number of codebooks. In this paper we have used six codebooks of size 1, 2, 4, 8, 16 and 32. We applied this technique over 20 PLP coefficients which have already been extracted from each voice sample, and which contains from 1 to 20 coefficients per person. This makes a total of 120 (6 * 20) extraction operations per person.

To train and validate our classifier, we used a technique of classification called LOPO, that is, we left out all the compressed frames of the PLP coefficients of one person to be used for validation as if it were an unseen person, and trained a classifier on the rest of the compressed frames of other people [6]. We used the LOPO technique of classification iteratively for each coefficient per person until all 20 coefficients per person for the six different codebooks size. In this paper, we used the SVMs classifier with its different types of kernels, i.e.; RBF, and Linear.

During the test section, we noticed that the obtained results when using a codebook size of 1 are not stable. Unlike the other codebook sizes, namely 2, 4, 8, 16 and 32, the compression of the PLP frames using a codebook size of 1, did not always give the same location of the centroids of the clusters forming the compressed PLP coefficients. Therefore, every time we redid the same test, we will not get the same classification results. To evaluate on how the results change, we used a test bed of 100 times. This test bed allows us to obtain the minimum, maximum and average value of the classification results for PD [19] [20] [21].

We made a test bed of 100 times, for the codebook size of 1, and 5 times for the codebook size of 2, 4, 8 and 16, and only one time for the codebook size of 32. As already mentioned, the obtained results using a codebook size higher than 1 are stable, nonetheless we did the test bed 5 times on the others to get an idea of the variation in execution time and to be sure that the results remained the same.

Based on our results, it is clear that using higher codebook size decreases the accuracy of classification as it is mentioned in Figure 3. It is clear from Table I that the time required for processing becomes longer.

The extracted PLP coefficients from each person contains in addition to the number of coefficients used, many frames with different values. The use of a large number of frames leads us to a diversity of results, often very close to the extracted values from other people (PD and healthy) [19]. This similarity of results between different people prevents making the correct diagnostic decisions [19]. By way to explanation; assuming that the frames are in the form of points distributed in space, increasing the number of frames leads to interference between these points [19]. Therefore, the task of the classifier, to find a hyper plane able to separate perfectly the two groups of people (namely patients with PD and healthy people), will be very difficult, if not impossible [19].

As can be seen in Table I, for a single test using a codebook size of 1, we need 78.71 seconds. Each time we increase the

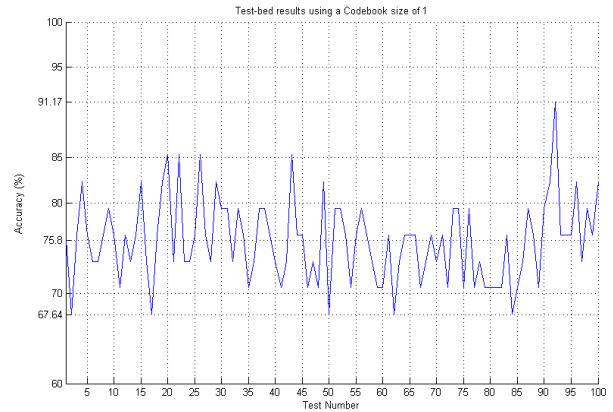


Fig 2: The test bed results using the codebook size of 1 with minimum, maximum and average classification accuracies

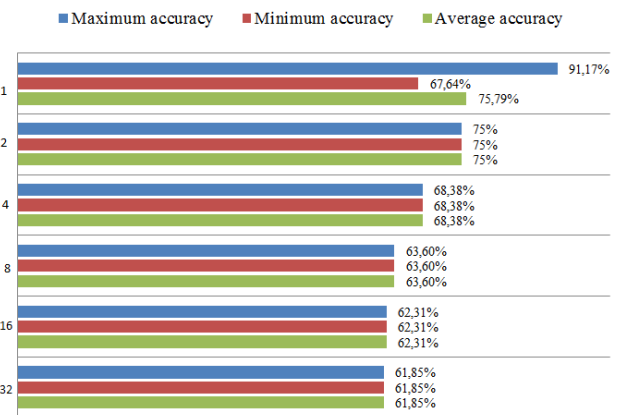


Fig 3: Classification results for different Codebook sizes (i.e. 1, 2, 4, 8, 16, and 32).

TABLE I
 EXECUTION TIME OF THE CLASSIFICATION PROGRAM FOR DIFFERENT SIZES OF THE CODEBOOK

Codebook sizes	Max time (second)	Min time (second)	Average time (second)
1	94.16	75.54	78.71
2	82.06	81.56	81.78
4	130.69	130.20	130.46
8	315.41	312.15	314.17
16	3.15e+03	3.13e+03	3.14e+03
32	4.94e+04	4.94e+04	4.94e+04

size of the codebook, the processing time for classification becomes longer. For a single test using a codebook size of 16 we need about 53 minutes and with a codebook size of 32 we need about 14 hours. For this size, it is not practical to apply a test bed if we already know that the results will remain the same even after 100 trials.

The test bed accuracy results using the codebook size of 1 are represented in Figure 2. A maximum classification accuracy of 91.17% was achieved using a codebook size of 1 as shown in Figure 3, by linear kernel SVMs. As seen from Figure 3, the best average classification accuracy of 75.8% was achieved using a codebook size of 1.

VI. CONCLUSION

A Dysarthria symptom associated with PD is a slow process whose early stages may go unobserved. To improve the evaluation of PD we collected a variety of voice samples from different people during the pronunciation of sustained vowel /a/. The extracted PLP coefficients from different people contain many frames which take maximum processing time in the classification section, and prevent making correct diagnosis. Therefore, we have compressed the extracted PLP coefficients using VQ with different codebook sizes.

After doing the tests we noticed that the obtained results using a codebook size of 1 were not stable. To evaluate on how the results change, we proceeded to a bench of 100 trials. The compression of the frames of the PLP coefficients using VQ with the codebook size of 1 has shown to be a good parameter for the detection of voice disorder in PD, showing an average classification accuracy of 75.8% and a maximum classification accuracy of 91.17%.

ACKNOWLEDGEMENT

The authors would like to thank M. Erdem Isenkul, Istanbul University. Thomas R. Przybeck and Daniel Wood, US Peace Corps Volunteers (Morocco 2013-2015). Dr. Robert Michael Hutchman, USA.

REFERENCES

- [1] Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J., & Ramig, L. O. (2009). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *Biomedical Engineering, IEEE Transactions on*, 56(4), 1015-1022.
- [2] Ishihara, L., and C. Brayne. "A systematic review of depression and mental illness preceding Parkinson's disease." *Acta Neurologica Scandinavica* 113.4 (2006): 211-220.
- [3] Jankovic, Joseph. "Parkinson's disease: clinical features and diagnosis." *Journal of Neurology, Neurosurgery & Psychiatry* 79.4 (2008): 368-376.
- [4] S. B. O'Sullivan, T. J. Schmitz, "Parkinson disease," *Physical Rehabilitation*, 5th ed. Philadelphia, PA, USA: F. A. Davis Company, 2007, pp. 856-894.2007, pp. 856-894.
- [5] Huse, Daniel M., et al. "Burden of illness in Parkinson's disease." *Movement disorders* 20.11 (2005): 1449-1454.
- [6] Sakar, Betül Erdogdu, et al. "Collection and Analysis of a Parkinson Speech Dataset With Multiple Types of Sound Recordings." *Biomedical and Health Informatics, IEEE Journal of* 17.4 (2013): 828-834.
- [7] Uma Rani, K., and Mallikarjun S. Holi. "Automatic detection of neurological disordered voices using mel cepstral coefficients and neural networks." *Point-of-Care Healthcare Technologies (PHT), 2013 IEEE. Bangalore, Indi*, pp. 76-79.
- [8] Little, Max A., et al. "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection." *BioMedical Engineering OnLine* 6.1 (2007): 23.
- [9] Rahn III, Douglas A., et al. "Phonatory impairment in Parkinson's disease: evidence from nonlinear dynamic analysis and perturbation analysis." *Journal of Voice* 21.1 (2007): 64-71.
- [10] Hermansky, Hynek. "Perceptual linear predictive (PLP) analysis of speech." *the Journal of the Acoustical Society of America* 87.4 (1990): 1738-1752.
- [11] R. Frail, JI. Godino-Llorente, N. Saenz-Lechon, V. Osma-Ruiz, C. Fredouille, "MFCC-based remote pathology detection on speech transmitted through the telephone channel," *Proc Biosignals*, Porto, 2009.
- [12] Jafari, A., "Classification of Parkinson's disease patients using nonlinear phonetic features and Mel-frequency cepstral analysis," *Biomedical Engineering: Applications, Basis and Communications* 25.04 (2013).
- [13] Ch. S. Kumar, P. R. Mallikarjuna, "Design of an automatic speaker recognition system using MFCC, Vector Quantization and LBG algorithm," *International Journal on Computer Science and Engineering*, Vol. 3, no. 8, 2011.
- [14] S. Young, G. Evermann, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, "The HTK Book (for HTK Version 3.4)," Copyright. 2001-2006, Cambridge University Engineering Department.
- [15] Makhoul, John, and Lynn Cosell. "LPCW: An LPC vocoder with linear predictive spectral warping." *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'76*. Vol. 1. IEEE, 1976.
- [16] Makhoul, John. "Spectral linear prediction: properties and applications." *Acoustics, Speech and Signal Processing, IEEE Transactions on* 23.3 (1975): 283-296.
- [17] Kapoor, Tripti, and R. K. Sharma. "Parkinson's disease Diagnosis using Mel-frequency Cepstral Coefficients and Vector Quantization." *International Journal of Computer Applications* 14.3 (2011), pp. 43-46.
- [18] J. Martinez, H. Perez, E. Escamilla, M. M. Suzuki, "Speaker recognition using mel frequency cepstral coefficients (MFCC) and Vector Quantization (VQ) techniques," *IEEE Electrical Communications and Computers*, Cholula, Puebla, Feb 2012, pp. 248-251.
- [19] Achraf BENBA, Abdelilah JILBAB and Ahmed HAMMOUCH. "Voice analysis for detecting persons with Parkinson's disease using MFCC and VQ." *The 2014 International Conference on Circuits, Systems and Signal Processing, September 23-25 2014, Saint Petersburg, Russia*. pp. 96-100.
- [20] Achraf BENBA, Abdelilah JILBAB and Ahmed HAMMOUCH. "Hybridization of best acoustic cues for detecting persons with Parkinson's disease," *2nd World conference on complex system (WCCS'14, November 10-12 2014, Agadir, Morocco, IEEE, 2014*.pp. 622-625.
- [21] Achraf Benba, Abdelilah Jilbab, Ahmed Hammouch, " Voiceprint analysis using Perceptual Linear Prediction and Support Vector Machines for detecting persons with Parkinson's disease", *the 3rd International Conference on Health Science and Biomedical Systems (HSBS '14), November 22-24 2014, Florence, Italie*, pp. 84-90.

Software for Reaction-Time Measurement and its Application for the Evaluation of Patient's Recovery after the Stroke

Mirjana Dejanović

Department of Physiology
Faculty of Medicine, University of Priština
Kosovska Mitrovica, Serbia
mirjana.dejanovic@gmail.com

Igor Dejanović

Department of Computing and Control Engineering
Faculty of Technical Sciences, University of Novi Sad
Novi Sad, Serbia
igord@uns.ac.rs

Abstract—The purpose of this study was to present a simple, cross-platform and easy to use software for reaction-time task and to find out the possibility of its application to the evaluation of patient's recovery after the ischemic stroke.

Keywords—*reaction-time; software; stroke;*

I. INTRODUCTION

The use of neuropsychological tests in patients after the stroke has become an obligatory part of the diagnostic protocol. Considering that the acute phase of a stroke greatly hinders the implementation of a comprehensive neuropsychological battery in most of these patients, emphasizes the importance of rapid tests for the assessment of cognitive status. Application of the method of reaction time is a quick and easy way to enable monitoring of mental processes that underlie a perception, attention, memory and action. The reaction time is the process that involves the receipt of information, its processing, decision-making and response-execution of motor acts [1]. The time interval from the moment of simple or complex stimuli presentation to the moment of the motor response reflects the speed of neurophysiological, cognitive and information processes that occurs as a response to the application of stimuli to the respondent sensory system.

Simple reaction time (SRT) includes reaction to the known stimulus and the same response is expected in subsequent attempts, thus the subject is able to preprogram the response move [2]. The response in the choice reaction time (CRT) is unknown until the appearance of the imperative stimulus, when the subject can plan, i.e. preprogram and initiate a response. In other words, the motor response was determined with at least 2 parameters where one of them can be changed between trials. In the choice reaction time, a response is not known in advance

and, therefore, advantages related to the anticipation cannot be used. Choice reaction time includes an analysis of stimuli and selection response, i.e. central cognitive processing [3].

A reaction time method test reflects the level of readiness of cognitive neural mechanisms, thus their use in the study of reaction speed in a variety of situations is tenable, both in healthy subjects and various diseases. Determining the value of simple and choice reaction time gives an insight into the functional state of the general reactivity of the individual in given time and circumstances.

Although a need for easily accessible good quality software for different types of reaction time is apparent we have found that the existing software is either proprietary and expensive or hard to use. Therefore, we have implemented easy to use cross-platform software for reaction time which will be presented in this paper. The software is in use at the Department for neurological disease at the Rehabilitation centre Novi Sad in Serbia and the Faculty of Medicine in Kosovska Mitrovica.

The aim of this work is to present our software for reaction time and its possibility for assessment of cognitive recovery of patients after the stroke.

II. RELATED WORK

There is not much reaction-time software that is freely available, easy to use and cross-platform.

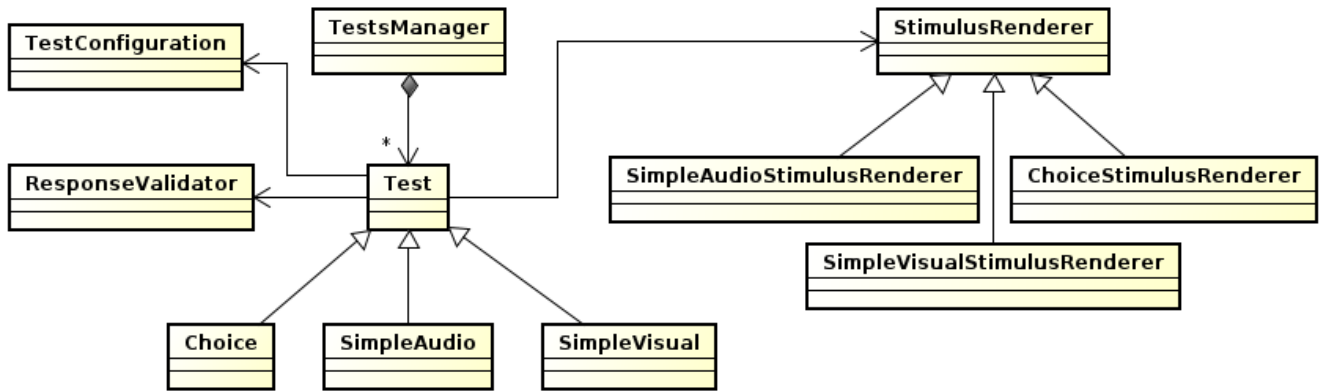


Fig. 1. Test and stimulus renderer class hierarchy

Commercially available software that can be used for reaction-time tasks, beside their price which might be a limiting factor to their use, are usually constrained to particular platform which hinders its use in the heterogeneous environments. DirectRT [4] is commercial reaction time software developed by Empirisoft. It claims to be very precise and easy to use. It is available on Windows platform only. Inquisit Lab [5] is also commercial software that can be configured to perform simple and choice reaction time tests.

There are freely available libraries and tools that can be used to help in building of reaction-time tests, but usually require at least some basic level of programming skills. PsychoPy [6] is a library and a toolset for Python programming language that can be used to build sophisticated psychological test, but require an understanding of Python language and PsychoPy application programming interface. OpenSesame [7] is another option where an experiment can be described using graphical user interface thus enabling less experienced users to built their experiments.

Tests of reaction time are proposed in the diagnosis of cerebral damage as indicators of its severity and extent [8]. In patients with stroke, simple and choice reaction time will be used for the assessment of cognitive status in the acute phase of stroke and also during the recovery [9].

Some studies show that simple and choice reaction times have a predictive role because faster choice reaction time in the acute phase of the stroke was significantly associated with better quality of life at 12 months [10]. The same authors found that simple and brief computerized assessment of attentional function in acute stroke is possible and is related to longer term attentional and cognitive performance. In [9, 11] authors have studied the efficacy of a computer-assisted reaction training on various attentional and cognitive functions in stroke patients with lateralized cortical lesions. All patients showed significant training effects for a number of attention functions, but not for vigilance, and there was no generalization of the training effects to more general cognitive functions [11]. However, training effects were more pronounced in the patients with lesions in the left hemisphere (or left brain-damaged group).

III. MATERIALS AND METHODS

The study included 60 patients (31 men and 29 women, mean age 59.7 ± 8.8 years, range 50-70 years) with acute brain infarction. Patients were recruited from the Department for neurological disease at the Rehabilitation centre Novi Sad Serbia. The control group consisted of 30, age and sex matched healthy volunteers (14 men and 16 women, mean age 58.5 ± 8.2 years, range 50-72 years). The protocol of the study was approved by the local Ethics Committee, in accordance with the principles of the Declaration of Helsinki. An informed consent was obtained from all patients, before they were enrolled into the study.

Simple and choice reaction time test was performed using a specially constructed computer program.

Simple reaction time test consisted of two blocks of tasks. The first being a visual stimuli task while the second was the audible. For simple reaction time task subjects were instructed to press Spacebar as fast as they can when the white circle was presented on the screen (visual stimuli), or the tone was played (audible stimuli). For each block stimuli were applied in a series of 30 repetitions where the time interval between two stimuli presentation varied between 1.5 and 3 seconds. For each subject a visual stimuli test was performed before the audible. From the 30 RT values for each serie (visual and audible) and average value was calculated and registered for the subject.

Choice reaction time test consisted of three blocks of tasks with 2, 4 or 6 choices each. Choices were given as digits 1-6 on the screen and subjects were instructed to press corresponding key on the keyboard when the digit is highlighted. As with simple reaction task, for each block, stimuli were applied in a series of 30 repetitions where the time interval between two stimuli presentations varied between 1.5 and 3 seconds.

IV. REACTION TIME SOFTWARE

Reaction is implemented in Java programming language and thus can run on every platform for which a Java runtime exists. This spans a wide variety of operating system environments.

Reaction currently supports two simple tests (visual and audible) and a choice reaction time test. From the graphical interface, an operator chooses a test to run. Tests are run in full-screen mode. Test session consists of one or more trial series. Each series consists of configured number of trials. At the beginning of each series, a user is presented with the short introduction where she is explained what she is supposed to do. Optionally, before each real series a practice series can be run. Practice series performs the same as the real one but the reaction data will not be collected and the number of trials will be smaller. In each trial, a subject must respond to the stimuli usually by pressing specific button on the keyboard. If the subject presses the button before the stimulus has appeared or presses the wrong button an error tone will be played and that particular trial will be marked as error. The number of erroneous trials is also an interesting datum to observe. At the end of each series collected data are displayed on the screen. At the end of the test session, an Excel file is written to the disk containing all collected reaction time data.



Fig. 2. Choice reaction time for 6 different stimuli

The main concept of *Reaction* is a test represented as *Test* class in Fig. 1. The tests instances are contained inside *TestManager*. Currently, tests are specified in Java by inheriting *Test* class but following versions will provide the means to specify tests using external textual description based on specially constructed language (i.e. Domain-Specific Language [12]). There are three types of test at the moment: *SimpleAudio*, *SimpleVisual*, and *Choice*.

Stimuli presentation is done by the instances of classes inheriting *StimulusRenderer* class. The job of stimuli renderers is to present stimulus to the user. The reaction is validated by the *ReactionValidator* instance.

SimpleAudio test type will play a tone of a specific frequency at random interval from the given minimum and maximum interval in milliseconds. *Reaction* validator will expect a press on spacebar to happen after the tone starts playing. The time that passes from the beginning of the tone to the key press will be recorded as RT.

SimpleVisual test type is similar to the *SimpleAudio* but will present a white circle at the center of the screen. Before a circle is presented a white + sign (i.e. fixation point) will be presented

to the user. This is needed for the user to better focus her attention while expecting a stimulus to appear.

Choice class implements a choice reaction time which displays a configurable number of digits from left to right centered horizontally and vertically on the screen. The stimulus, in this case, will be a random highlight of one of the digit as shown in Fig. 2. Expected response is pressing a corresponding digit on the keyboard. *Choice* can be configured with 2-10 digits to appear, but it is usually configured for 2, 4, 6 digits. Recorded reaction time will be a time that passes from the highlight to the press on the right digit on the keyboard. If the user presses the wrong key the trial will be recorded as an error.

TABLE I. RESULTS OF CHOICE REACTION TIME TEST FOR PATIENTS AND CONTROL GROUP AT ACUTE PHASE OF THE STROKE

CRT	Patients [ms]	Controls [ms]	t	p
2	710.2	486.4	-4.33	<0.01
4	985	705	-4.23	<0.01
6	1305	870	-5.25	<0.01

The results obtained using choice reaction time test, presented in Table I, shows that the RT values for patients with ischemic stroke at the beginning of the study greater than the values of the subjects in the control group and that difference is statistically significant. At second registration, three months after the first registration a statistically significant recovery of RT values have been observed.

V. CONCLUSION

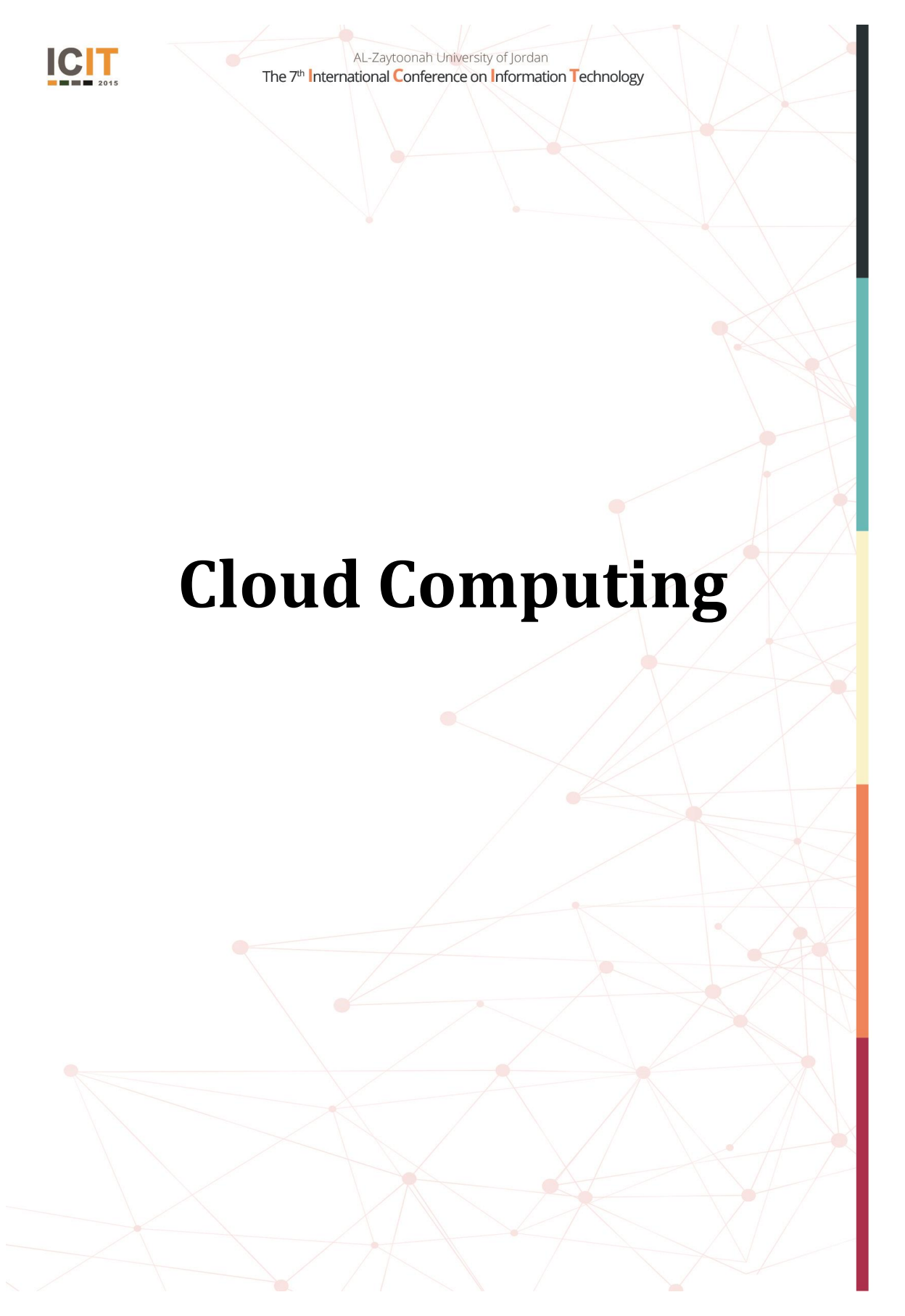
We have described cross-platform software for performing simple and choice reaction-time tasks which we used to evaluate patient’s recovery after the ischemic stroke. We have found that using a computer-based reaction time test is a quick and easy way to assess the cognitive abilities of patients after an ischemic stroke. This may be a significant aid in the forecast of their recovery.

REFERENCES

- [1] Pachela R.G., “The interpretation of reaction time in information-processing research”, In Human information processing – tutorials in performance and cognition. Ed. Barry H. Kantowitz. New York, 1978.
- [2] Kohfeld D.L., “Simple reaction as a function of stimulus intensity in decibels of light and sound”, *Journal of Experimental Psychology*, 1971;88:251-257.
- [3] Tenenbaum G., Yuval R., Elbaz G., Bar-Eli M. and Weinberg R., “The relationship between cognitive characteristics and decision making”, *Can J Appl Physiol*, 1993, Mar; 18 (1):48-62.
- [4] “Direct RT”, Empirisoft, <http://www.empirisoft.com/directrt.aspx>, online, accessed January, 24. 2015.
- [5] “Inquisit Lab”, Millisecond, <http://www.millisecond.com/>, online, accessed January, 24. 2015.
- [6] Peirce, JW, “PsychoPy - Psychophysics software in Python”, *J. Neurosci Methods*, 162(1-2):8-13. 2007.
- [7] Mathôt, S., Schreij, D., and Theeuwes, J. “OpenSesame: An open-source, graphical experiment builder for the social sciences”. *Behavior*

- Research Methods, 44(2), 314-324. doi:10.3758/s13428-011-0168-7. 2012.
- [8] Dee H. L. and Van Allen M. W., "Psychomotor testing as an aid in the recognition of cerebral lesions". *Neurolog.v* 22, 845-848, 1982.
- [9] Cumming TB1, Brodtmann A, Darby D. and Bernhardt J., "Cutting a long story short: reaction times in acute stroke are associated with longer term cognitive outcomes". *J Neurol Sci.* 2012 Nov 15;322(1-2):102-6.
- [10] Cumming TB, Brodtmann A, Darby D, Bernhardt J., "The importance of cognition to quality of life after stroke", *J Psychosom Res.* 2014 Nov;77(5):374-9.
- [11] Walter Sturm and Klaus Willmes, "Efficacy of a reaction training on various attentional and cognitive functions in stroke patients", *Neuropsychological Rehabilitation: An International Journal* Volume 1, Issue 4, 1991 pages 259-280
- [12] Fowler, M., "Domain-Specific Languages", Addison-Wesley Professional, 2010.

Cloud Computing



SwiftEnc: Hybrid Cryptosystem with Hash-Based Dynamic Key Encryption

Yasir S. Alagl, El-Sayed M. El-Alfy
College of Computer Sciences and Engineering
King Fahd University of Petroleum and Minerals
Dhahran 31261, Saudi Arabia
{g200720290, alfy}@kfupm.edu.sa

Abstract—With the emerging need to store massive data in cyberspace and cross platforms, whether in local file systems or cloud-based services, certain security requirements must be met to efficiently protect confidentiality and privacy and manage the large number of keys and access policies. Most of the current encryption standards emphasize one of two trade-off factors: speed of encryption versus ease of key management. Though asymmetric-key encryption does not require the sender and receiver to share a common secret similar to symmetric-key encryption, the cost of the mathematical computations may be unaffordable. In this paper, we first review the state-of-the-art of hybrid cryptosystems. Then, we propose a novel scheme for lightweight encryption of bulk data based on recursive cryptographic hashes and dynamic keys. The effectiveness of the proposed scheme is demonstrated on three files having different sizes, types and contents.

Keywords—data security; bulk data encryption; cryptographic hashing; hybrid cryptosystems; dynamic keys; password-based key derivation; security vault.

I. INTRODUCTION

As the Internet grows in size and number of users, new technologies and applications inevitably emerge to comply with such growth and to satisfy various demands of users. Along with this growing and collaborative environment, the dependability on multiple technological platforms that serve as tools in accommodating many aspects of day-to-day tasks also increases. However, with the tremendous benefits these services provide comes the struggle of protecting users' sensitive data and files within underlying cross-platform systems. The ability to backup, share and synchronize files and folders is becoming crucial over time. The trend to store large volumes of various types of data and files securely is tempting due to durability, portability, flexibility and ease of share, and resistance to threats.

At the heart of security defense mechanisms, encryption arises to protect the confidentiality of valuable data from unauthorized access by programs and individuals [1]. However, at relatively high computational costs, encryption is usually delegated to other parties or skipped in total, thus, exposing the value of an asset to threats [2]. For example, Dropbox, which is a widely-used cloud-based service for hosting files, has been criticized for its weak protection of user's privacy since its first release in September 2008. Lately, similar to a competitive service known as SpiderOak, Dropbox allowed its customers to encrypt their files on the server using the Advanced Encryption Standard (AES) with 256-bit key. In contrast, SpiderOak stores an encrypted version of the decryption key as well in a manner that even the company's employers will not be able to decrypt these files without knowing the customer's password [3]. However, if an intruder managed to get that key, all files can be decrypted.

Generally, cryptosystems fall under two broad categories: symmetric and asymmetric [5]. Although symmetric-key encryption is proven to be relatively faster than asymmetric-key encryption [4], it suffers from two issues. First, it requires sharing a key between the encryption and decryption entities, which might be in different systems. Second, it requires a large number of unique shared keys. Consider a group of N members who engage in an exchange process of a valuable asset T times. Furthermore, consider that each exchange requires an asset to be encrypted with a uniquely generated symmetric key prior to exchanging it. Each member should encrypt a given asset $(N - 1) \times T$ times, in addition to sharing $(N - 1) \times T$ symmetric keys through other secure channels. Moreover, consider having a pool of assets all of which require exchange. The reader can notice the exponential growth in the number of keys and the overhead of sharing them. Examples of the popular symmetric-key encryption standards are Blowfish, International Data Encryption Algorithm (IDEA), Data Encryption Standard (DES), and Advanced Encryption Standard (AES).

Asymmetric encryption, on the other hand, doesn't require the disposal of keys upon each exchange, due to the concealment of the private key. Consider the previous scenario, however, with asymmetric encryption as a requirement for assets exchange. Each member in the group announces his own public key that should be used prior to commencing an exchange with him. This key can still be used with every subsequent exchange resulting in eliminating the overhead of key exchange and the generation of keys. The security of asymmetric encryption depends on the intractability of the discrete logarithm problem and hence comes with higher costs for the encryption and decryption process, i.e. it is relatively slower to encrypt bulk files. Examples of popular asymmetric-key encryption are RSA and ElGamal cryptosystems [5].

Here comes the need of an algorithm that combines the merits of the two categories into what is known as hybrid cryptosystems [6]. The goal is to use the public/private key pairs, but maintain superior performance than that of asymmetric encryption. PGP, GnuPG and OpenPGP are examples of the popular hybrid cryptosystems [7]. Another example is a proprietary standard used by Microsoft for Encrypting File System (EFS) since the release of Windows NT Version 3.0.

In this paper, we introduce SwiftEnc, a lightweight hybrid scheme that can be used effectively to encrypt bulk data. SwiftEnc is a hash-based obfuscation algorithm that uses variable-length dynamic key computed based on the file to be encrypted. It also uses an existing asymmetric encryption algorithm to encrypt. The goal is to produce a cipher text in relatively faster time than those of asymmetric algorithms. Based on this scheme, a prototype of a security vault is developed for managing keys in a central store such as Windows Registry.

The rest of the paper is organized as follows. Section II reviews related work. Subsequently, Section III presents the proposed scheme, SwiftEnc algorithm, and describes each of its components in details. We then provide benchmarks comparing the proposed algorithm to existing encryption algorithms in Section IV. Finally, the paper conclusion is given in Section V.

II. RELATED WORK

There have been many trails in both the academic and industry sectors to produce fast encryption algorithms. However, each within its own domain, there hasn't been much work on a general-purpose algorithm that encrypts any file with adequate performance. In this section, we shed the light on some of the recent work that has been made on fast encryption. Presumably, AES is considered the fastest accepted standard of an encryption algorithm worldwide [8]. However, it may not be suitable for very constrained environments and still more improvements are needed [9].

In [10], Wang et al. discussed the use of chaos-based fast image encryption algorithm for image encryption. They proposed combining the scanning process of an image on both stages of permutation and diffusion into one, thus reducing the time required for scanning dramatically. They partitioned the image into blocks of pixels and shuffled the blocks using spatiotemporal chaos and diffused them to change the pixel value at the same time. They also presented an efficient method for pseudo-random generation that is used within their algorithm [11].

In [12], Verkhovsky explained the nature of encryption using Gaussians that belong to complex numbers family. He proposed a new algorithm that finds all cubic roots of Gaussian integers. The algorithm introduces some constraints with regards to encryption time. However, decryption is substantially slower than encryption and hence it only fits applications where only the sender has limited time.

In [13], Hohenberger and Waters introduced an Attribute-Based Encryption (ABE) algorithm with fast decryption. ABE is an expansion of public-key encryption that allows users to encrypt and decrypt messages based on their attributes. However, the complexity of decryption increases as more attributes are utilized. The proposed ABE scheme allows a cipher text to be decrypted with constant number of pairing, specifically 2 pairings, by increasing the private key size.

In [14], LAE is described as a high-speed software block cipher that competes with AES on all standard platforms such as Intel, AMD and ColdFire. LAE works with 128-bit block size and similar key sizes to those of AES, i.e. 128, 192, and 256. It's shown that LAE is faster than AES due to the use of ARX operations (modular Addition, bitwise Rotation, and bitwise XOR) which are supported on most 32-bit and 64-bit platforms. Moreover, LAE rounds are all the same without requiring special end round. The authors also showed that LAE is secure against existing attacks.

Among the attempts to develop encryption algorithms with low implementation complexity comes a promising class of lightweight techniques [15], [16], [17], [18]. For instance, PRESENT is a lightweight block cipher that has been shown to be 2.5 times faster than AES [9].

The concept of dynamic keys or sequence of one-time symmetric cryptographic keys is described and analyzed in [19]. Based on this analysis, the advantages of dynamic keys are revealed in terms of security and efficiency. In essence, if the hacker is able to expose one message, the other messages remain secure. Lastly, in [20] and [21] some trials were made to accelerate the encryption process by the use of Graphical Processing Unit (GPU). However, these trials targeted High Performance Database Management System (DBMS). In [22], the use of GPUs was also noted to accelerate homomorphic encryption.

Some systems and platforms have developed to provide solutions for big data and to establish secure vault for the increased number of keys, certificates and policies. Examples of these systems are the IBM InfoSphere [24], Oracle TDE [25], Microsoft TDE [26], and Volumetric Data Security products [27].

III. PROPOSED CRYPTOSYSTEM

In this section, we provide details on the proposed scheme, SwiftEnc. The implementation of SwiftEnc is aimed to be flexible and lightweight. Any available encryption algorithms can be included as long as they meet the requirements of SwiftEnc. The proposed scheme starts by acquiring a secret phrase (passphrase) from the user. This passphrase is used to generate a pair of public and private keys for the chosen asymmetric encryption algorithm. The private key can be discarded at this point while the public key should be stored. The user then selects a file to be encrypted and generates key material or key seed, h_0 , from the file itself and some secure pseudo-random numbers. Once the key material is generated, it is passed through a sequence of hashing. To increase the security by maximizing the entropy, the input to each hashing

step is output from the previous step XORed with a counter. The process stops once we acquire a bulk key, K_s , that has equal length to the file we intend to encrypt.

To encrypt the file, a simple operation similar to stream cipher is then used. In our case, we XOR the key, K_s , with the file to generate an obfuscated secure output file that can be shared over insecure medium or stored locally. Meanwhile, the public key that was generated from the user supplied passphrase is used to encrypt the initial seed, h_0 , and store it with the obfuscated file. Figure 1 shows an outline for the process of encryption in SwiftEnc. The subsequent subsections provide more details on our implementation of the proposed SwiftEnc cryptosystem.

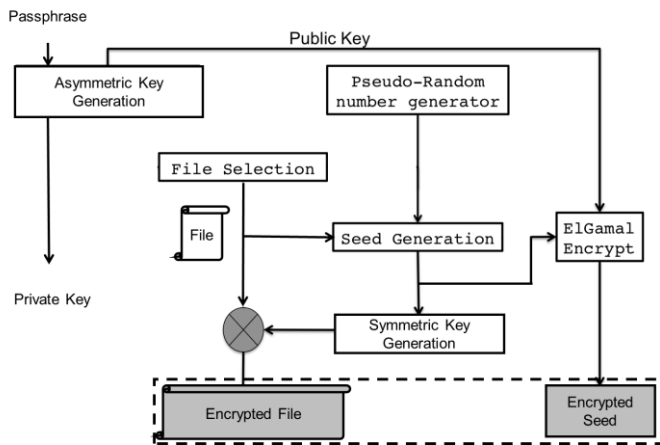


Fig. 1. Outline of the SwiftEnc process.

A. Key Generation and Management

Asymmetric encryption is used to protect the file encryption key seed, h_0 . Our choice for asymmetric key generation and sharing in SwiftEnc is ElGamal public-key cryptosystem [23]. However, should the implementer of SwiftEnc make use of other asymmetric encryption algorithm, the general scheme still holds. For instance, using RSA requires the use of two keys: public key and private key. While the private key should be kept safe and private by its owner from unauthorized access (as the name suggests), the public key does not. Assuming a single platform on which the user intends to encrypt his assets or files for his own use, the public key can be kept in his home directory or in any sort of non-protected data store, e.g. Windows Registry. In a scenario where a group of members communicate securely back and forth, each member's public key should be announced within the group together with a certificate to authenticate the validity of the public key.

To avoid the need to a trusted third party to issue a certificate, we use ElGamal algorithm for key generation and sharing. This algorithm is based on Diffie-Hellman key exchange and uses two keys at each of the sender and the

receiver sides. These keys are generated in such a way to allow them to share a session key. For example, assume A is the sender and B is the receiver. Then, A should have $K_{prv,A}$ and $K_{pub,A}$, and B has $K_{prv,B}$ and $K_{pub,B}$. The receiver, B , starts by defining a cyclic group G of order p , where p is a large prime number. This cyclic group has a generator g . B then selects a private key $K_{prv,B} < p - 1$ randomly from G and calculates a corresponding public key $K_{pub,B}$ as follows:

$$K_{pub,B} = g^{K_{prv,B}} \text{ mod } p \quad (1)$$

B announces the tuple $(K_{pub,B}, g, p)$ or stores it in a shared folder. If A wants to securely send a file to B , it should obtain the tuple $(K_{pub,B}, g, p)$ and selects a private key $K_{prv,A}$ from the group G generated by (g, p) . Then, A calculates an ephemeral public key $K_{pub,A}$ as follows:

$$K_{pub,A} = g^{K_{prv,A}} \text{ mod } p \quad (2)$$

It also calculates a shared key, K_m , to be used for encrypting the file encryption key seed, h_0 . The calculation of K_m is as follows:

$$K_m = (K_{pub,B})^{K_{prv,A}} \text{ mod } p \quad (3)$$

K_m will be used to encrypt the message using ElGamal encryption algorithm and the encrypted message together with $K_{pub,A}$ will be submitted to the receiver. The decryption will be performed using inverse operation.

As an alternative approach in SwiftEnc, we used SHA-512 to hash the passphrase and the result x is identified as our private key. The passphrase could be of any length, complexity, and combination of characters' groups, e.g. uppercase, lowercase, special characters, numbers, etc. The use of a passphrase introduces usability rather than remembering a random number. The passphrase can be fixed for all files or can be changed for each file. In our case, we made it fixed for all files in the vault. The passphrase goes through a one-way hashing function such as MD5 or SHA-512 to produce a fixed-length hash string then use the first 128 bits or 512 bits, for example, as $K_{prv,B} = x$. Password-based key derivation is common in practice and industry standards such as PKCS and OpenPGP. In [28], a framework for the design and analysis of password-based key derivation functions (KDFs) is provided.

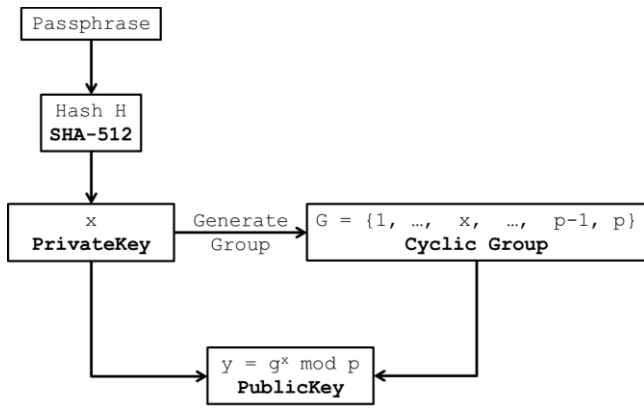


Fig. 2. SwiftEnc public/private key generation using ElGamal.

Afterthat, G is chosen such that: $x \in G$, x can be generated by g to some order $1 < x < p - 1$ where p is a prime number. In our implementation, we used the `BigInteger.probablePrime()` method in Java to generate a value for p . Once these conditions are met, we can identify our public key as per ElGamal and discard the private key, x , completely. Hence, avoid the overhead of storing an encrypted version of the private key; it can be generated whenever needed from hashing the passphrase. We can also pass x to the hash function, for the second time, and produce x_1 which can be used to check the validity of the entered passphrase in later operation, i.e. decryption. Figure 2 illustrates this process for public/private key generation.

One important concept to note here is that SwiftEnc doesn't use the public/private key pair for encrypting/decrypting assets. Indeed, the passphrase and generation of keys don't account for the confidentiality of the asset by any factor. However, the encryption/decryption depends on the asset itself as will be discussed in the subsequent sections.

B. Seed Generation

In SwiftEnc context, the seed refers to the string of characters that will be used to generate a symmetric key that, in turn, will be used to obfuscate the asset which the user intends to encrypt. However, this seed varies in length and value per each file. The seed is the secret that we want to insure that it's properly protected, as obtaining the seed reveals the confidentiality of the asset as we will see in Section III-D.

Since every file will have its own unique seed, the seed has to be stored along with the protected asset, however, in a confidential format. We will see in subsequent sections that the seed is necessary to decrypt the asset and return the file to its original state. The implementation of SwiftEnc can use any seed generation algorithm to associate a seed to a file under the following conditions: (a) The algorithm guarantees a sufficient degree of pseudo-randomness, and (b) The algorithm acquires very low probability of collision. In SwiftEnc, we create the seed from the first block of the file to be encrypted as indicated in Algorithm 1.

```

Data: File to encrypt (FE), Initial seed size (SS)
Result: Initial seed (h0)
FS = FE.getSize();
count = 0;
while count < min(SS, FS) do
    h0[count] = FE.getNextByte();
    h0[count+1] = SecureRandom();
    count += 2;
end
while count < SS do
    h0[count] = SecureRandom();
    count++;
end
return h0;
    
```

Algorithm 1. Seed creation algorithm.

C. File Obfuscation

Once the seed is generated for the perspective asset that we intend to protect, the Symmetric Key Generation process and file obfuscation can start immediately. Obfuscation is the core of SwiftEnc on which the asset's data are being randomly scrambled to generate an encrypted file. Moreover, this operation occurs with minimal processing power and fast timing, hence the term Swift. To assure that SwiftEnc accommodates larger file sizes, we use buffered streams to process the file sequentially.

We generate a key from the seed by recursively hashing it. Since SwiftEnc is using SHA-512, the first 512 bits (64 bytes) of the key will be the hash of the initial seed h_0 . The following 64 bytes will contain the hash of the resulting hash from the previous step, and so on. We repeat this operation until we reach a key equal in length to the first 64 bytes multiple that is larger than the file size. Next, we perform a regular XOR operation between each byte of the asset and the key and send/store the result as our encrypted file. The use of XOR with the hash gives SwiftEnc the low processing power and better performance over other encryption algorithms, however, we haven't discussed what gives it a confidentiality level. Algorithms 2 and 3 demonstrate these processes. The illustration of the prototype operation is depicted in Fig. 3. The XOR operation is reversible in nature. So, we can use this property to decrypt the file and retrieve the original cleartext file. By only having the encrypted file and the seed, we can generate the same key by hashing the seed recursively and XORing it with the file, thus, revealing our file back.

```

Data: File to encrypt ( $FE$ ), Initial Seed ( $h_0$ )
Result: Symmetric encryption key ( $K_s$ )
 $FS = FE.getSize();$ 
 $SS = h_0.getSize();$ 
 $count = ceil(FS / SS);$ 
 $key_0 = hash(h_0);$  //  $key_0$  subscript means block
 $K_s = key_0[0..SS];$  take the first  $SS$  bytes
 $i = 1;$ 
while  $i < count$  do
     $key_i = hash(key_{i-1} \oplus i);$ 
     $K_s = K_s || key_i[0..SS];$  // concatenation
     $i ++;$ 
end
return  $K_s;$ 
    
```

Algorithm 2. Symmetric key generation for file encryption.

```

Data: File to encrypt ( $FE$ ), Key ( $K_s$ )
Result: Encrypted file ( $EF$ )
 $FS = FE.getSize();$ 
 $i = 0;$ 
while  $i < FS$  do
     $EF[i] = FE[i] \oplus K_s[i];$ 
     $i ++;$ 
end
    
```

Algorithm 3. File obfuscation

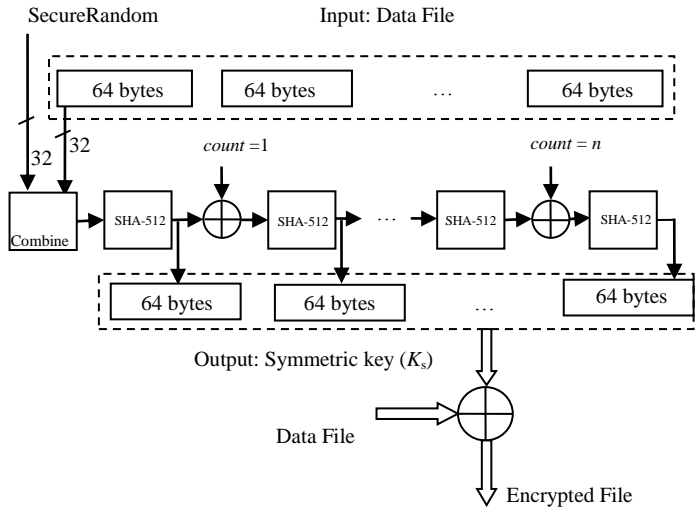


Fig. 3. Main steps for generating symmetric key and encrypting the data file.

D. Seed Cryptography

As we have seen, the seed is generated by extracting its value from the file and a pseudo-random number generator. The seed is also used to create a key that is equal in length to the length of the file. Once we obtain the key, we can encrypt our asset immediately and discard the key completely. However, obtaining the seed compromises the security of the system and redeems the asset unsecured. Once an unauthorized entity obtains the seed, our asset is no longer protected.

To thwart against such threat, the owner of the asset should provide a layer of protection over the seed. SwiftEnc ensures that this layer is implemented by encrypting the seed using any well-known Asymmetric Encryption algorithm; in our case we have chosen ElGamal as discussed above using K_m from Eq. (3):

$$h'_0 = h_0 \cdot K_m \pmod p \quad (4)$$

The encrypted seed, h'_0 , should be stored next to the file, appending/pre-appending it to the file, or in a data-house that could link it to the file. Upon decryption, we should retrieve the seed with respect to the file, decrypt it using ElGamal, then initiate the de-obfuscation as stated in Section III-C. To decrypt the seed, ElGamal has to calculate K_m at the receiver then use its inverse in the cyclic group G to decrypt the seed:

$$K_m = (K_{pub,A})^{K_{priv,B}} \pmod p \quad (5)$$

$$h_0 = h'_0 \cdot K_m^{-1} \pmod p \quad (6)$$

IV. EVALUATIONS

We have developed the algorithm described above and built a prototype for a security vault as a central location for managing encrypted files and passwords. Figure 4 shows part of the user interface for the main menu and the security vault. We report some empirical experiments to benchmark SwiftEnc with another password-based hybrid encryption algorithm (Rijndael-RSA) [29]. Rijndael-RSA encrypts and decrypts using 256-bit Rijndael key where the key is encrypted using 1024-bit RSA key, which is password-encrypted. All implementations were conducted in Java and experiments were run on the same machine using the specifications shown in Table I.

TABLE I. EXPERIMENTS SPECIFICATION

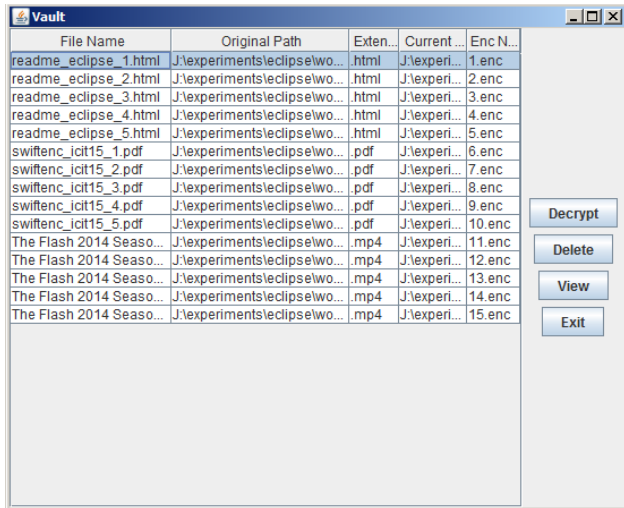
OS	64 bit Windows 7 Professional
Processor	Intel Core i5-33M CPU 2.7GHz
Memory	4 GB
Implementation	Java 1.7

The algorithms are tested on three files of different sizes and content types. The first file is the readme file that comes with eclipse and contains HTML. The second file is the PDF of an initial version of this paper. The third file is MP4 file

corresponding to “The Flash 2014 Season 1 Episode 01”. The performance measures are reported in terms of: average time and speed. The time includes I/O reading and writing times, the key generation, encryption and decryption. The speed is calculated as size in MB divided by time in seconds. Table II illustrates the average times in seconds for five runs as well as the speed.



(a) Main menu interface



(b) Vault interface

Fig. 4. Screenshot a security vault prototype based on SwiftEnc for encryption and decryption.

TABLE II. COMPARISON OF AVERAGE OVERALL TIME (SEC) AND SPEED (MB/SEC) APPROX TO 4 DECIMAL DIGITS

Input File		SwiftEnc		Rijndael-RSA	
Size (MB)	Type	Time	Speed	Time	Speed
0.1	HTM	0.2444	0.4092	2.9204	0.0342
0.773	PDF	0.6598	1.1716	14.6696	0.0527
272	MP4	187.7854	1.4485	4925.0770	0.0552

V. CONCLUSION

This paper discussed the trade-offs of encryption algorithms and how they can impose a barrier on the value of assets due to their relatively high processing time. We introduced a new hybrid algorithm, SwiftEnc, and security vault prototype, that can be used to overcome this barrier and allow for rapid encryption with low processing power. The vault prototype provides a central local store for securely managing keys and encrypted files. The framework can be customized with different cryptographic functions to accommodate various security standards enforced by an organization. SwiftEnc showed better performance when

compared to another algorithm. When used for communication over the Internet, message exchanges between the sender and the receiver can also include timestamp and nonce to counter replay attacks.

ACKNOWLEDGMENT

The authors would like to acknowledge the support provided by King Fahd University of Petroleum & Minerals (KFUPM) during this work.

REFERENCES

- [1] D. R. Stinson, *Cryptography: Theory and Practice*. CRC Press, 2005.
- [2] A. Nadeem and M. Y. Javed, “A performance comparison of data encryption algorithms,” in *Proc. IEEE International Conference on Information and Communication Technologies*, 2005, pp. 84–89.
- [3] S. Latha, K. Raju, and S. Santhi, “Overview of dropbox encryption in cloud computing,” *Transactions on Engineering and Sciences*, vol. 2, no. 3, pp. 27–32, 2014.
- [4] A. Al-Hasib and A. Haque, “A comparative study of the performance and security issues of AES and RSA cryptography,” in *Proc. Third International Conference on Convergence and Hybrid Information Technology, ICCIT '08*, vol. 2, Nov 2008, pp. 505–510.
- [5] A. J. Menezes, P. C. Van Oorschot, and S. A. Vanstone, *Handbook of Applied Cryptography*. CRC Press, 2010.
- [6] A. W. Dent, “Hybrid cryptography,” Information Security Group, University of London, Tech. Rep., 2005.
- [7] The International PGP Home Page. [Online]. Available: <http://www.pgp.org/>
- [8] F. P. Miller, A. F. Vandome, and J. McBrewhster, *Advanced Encryption Standard*. Alpha Press, 2009.
- [9] A. Bogdanov, L. R. Knudsen, G. Leander, C. Paar, A. Poschmann, M. J. Robshaw, Y. Seurin, and C. Vikkelsoe, “PRESENT: An ultralightweight block cipher,” in *Cryptographic Hardware and Embedded Systems - CHES 2007, Lecture Notes in Computer Science*, vol. 4727. Springer, 2007, pp. 450–466.
- [10] Y. Wang, K.-W. Wong, X. Liao, and G. Chen, “A new chaos-based fast image encryption algorithm,” *Applied Soft Computing*, vol. 11, no. 1, pp. 514–522, 2011.
- [11] M. Bellare, A. Desai, E. Jorjipii, and P. Rogaway, “A concrete security treatment of symmetric encryption,” in *Proc. 38th Annual Symposium on Foundations of Computer Science*, 1997, pp. 394–403.
- [12] B. Verkhovsky, “Cubic root extractors of gaussian integers and their application in fast encryption for time-constrained secure communication,” *Int. J. of Communications, Network and System Sciences*, vol. 4, pp. 197–204, 2011.
- [13] S. Hohenberger and B. Waters, “Attribute-based encryption with fast decryption,” in *Public-Key Cryptography–PKC 2013, Lecture Notes in Computer Science*. Springer, 2013, vol. 7778, pp. 162–179.
- [14] D. Hong, J.-K. Lee, D.-C. Kim, D. Kwon, K. H. Ryu, and D.-G. Lee, “LEA: A 128-bit block cipher for fast encryption on common processors,” in *Information Security Applications*. Springer, 2014, pp. 3–27.
- [15] T. Eisenbarth, S. Kumar, C. Paar, A. Poschmann, and L. Uhsadel, “A survey of lightweight-cryptography implementations,” *IEEE Design & Test of Computers*, vol. 24, no. 6, pp. 522–533, 2007.
- [16] B. Adida, S. Hohenberger, and R. L. Rivest, “Lightweight encryption for email,” in *USENIX Steps to Reducing Unwanted Traffic on the Internet Workshop (SRUTI)*, 2005.
- [17] E. Choo, J. Lee, H. Lee, and G. Nam, “SRMT: A lightweight encryption scheme for secure real-time multimedia transmission,” in *Proc. International Conference on Multimedia and Ubiquitous Engineering*, 2007, pp. 60–65.

- [18] D. Engel and A. Uhl, "Lightweight JPEG2000 encryption with anisotropic wavelet packets," in *Proc. IEEE International Conference on Multimedia and Expo, ICME '06*, 2006, pp. 2177–2180.
- [19] H. H. Ngo, X. Wu, P. D. Le, C. Wilson, and B. Srinivasan, "Dynamic key cryptography and applications," *International Journal of Network Security*, vol. 10, no. 3, pp. 161–174, 2010.
- [20] H. Jo, S.-T. Hong, J.-W. Chang, and D. H. Choi, "Data encryption on gpu for high-performance database systems," *Procedia Computer Science*, vol. 19, pp. 147–154, 2013.
- [21] —, "Offloading data encryption to GPU in database systems," *The Journal of Supercomputing*, pp. 1–20, 2014.
- [22] W. Wang, Y. Hu, L. Chen, X. Huang, and B. Sunar, "Accelerating fully homomorphic encryption using GPU," in *Proc. IEEE Conference on High Performance Extreme Computing (HPEC)*, 2012, pp. 1–5.
- [23] T. ElGamal, "A public key cryptosystem and a signature scheme based on discrete logarithms," in *Advances in Cryptology*. Springer, 1985, pp. 10–18.
- [24] <http://www-01.ibm.com/software/data/infosphere/>
- [25] <http://www.oracle.com/technetwork/database/options/advanced-security/index-099011.html>
- [26] <https://msdn.microsoft.com/en-us/library/bb934049.aspx>
- [27] <http://www.vormetric.com/>
- [28] F. F. Yao, and Y. L. Yin, "Design and analysis of password-based key derivation functions," in *Topics in Cryptology*, Springer, 2005, pp. 245–261.
- [29] J. Garms and D. Somerfield, *Professional Java Security*. Wrox. 2001.

Proposed Evaluation Criteria for Selecting Appropriate Cloud Based On-Demand CRM for SMEs.

Musarat Hasan Mujawar

Middlesex University, London, United Kingdom.
Capgemini Consulting, India.

musarat.hasanmujawar@gmail.com, musarat.mujawar@capgemini.com

Abstract—On-Demand CRM is of particular significance to small and medium-sized organizations as they are faced with budget and time restrictions. However, the failure rate of CRM in organizations is still considerably high. The CRM On-Demand market is saturated with multiple vendors offering similar solutions. Due to technological advancements, SME managers tend to not understand thoroughly the key drivers of On-Demand CRM solution. Some managers have left much of the understanding of the technical artifacts to the cloud vendors. There exists a research gap, which renders technical variables not part of the management decision-making activities. This research aims to address this gap in the CRM domain and focus on the technical aspects of evaluating On-Demand CRM in a bid to aid managers in decision-making. The design science research approach is adopted. Quantitative data analysis methods were used to test and validate the proposed evaluation criteria. The results indicate that from a technical quality perspective, there are various non-functional attributes that form a part of the CRM evaluation process in the cloud. This framework can be improved upon further by studying the functional, process and people factors that affect the buying decision of CRM.

Keywords— *On-Demand; CRM; evaluation; cloud.*

I. INTRODUCTION

CRM (Customer Relationship Management) is a technology-based and technology-integrated business process management strategy for applications marketing, sales and service that maximizes relationships between customer and organizations [4]. Cloud computing combined with On-Demand services over SAAS (Software-as-a-Service) architecture has penetrated CRM systems, too. Customers can access applications and data from a “cloud” anywhere in the world on-demand [1]. This has convinced many organizations to buy SAAS-based solutions irrespective of their geographic location. The On-Demand CRM market comprised 32% in 2011 and 39% in 2012 of the total CRM market [2]. Hence, On-Demand CRM cloud-based SAAS applications for businesses are seen as a growing trend among organizations, especially SMEs (Small and Medium Enterprises). SMEs are defined as small and medium sized enterprises with a number of employees not greater than 250 and an annual turnover not greater than EUR 50 million [3]. Another point worth considering is that customer retention is also becoming increasingly important to SMEs because of their limited resources. Hence, SMEs, with their restricted budgets and deadlines look at cloud-based hosted CRM solutions as an attractive option. This is highly influenced by the pricing model of cloud-based SAAS CRM, which is valued on a per user per month basis [4]. The customer only has to pay for the usage of the service, which is much cheaper when compared to

traditional On-Premise solutions, which entail high installation costs upfront [5]. However, for the sake of low costs and quick installation, the SME can be tricked into buying a solution that may not meet the business needs. This may result in loss of customers, which is unaffordable and a huge risk for the SME. Along with insufficient knowledge of On-Demand models of CRM and technological factors, AMR Research found that 47% of companies surveyed reported serious challenges with end user adoption [6].

The evaluation of On-Demand CRM selection is often confused with cloud service evaluation only, thereby neglecting other factors that can be CRM specific, as well as a mix of CRM and cloud technology. This requires managers to narrow their focus to On-Demand CRM systems within the cloud.

II. RELATED CONCEPTS

This section provides an overview of the concepts like cloud computing, cloud architecture and On-Demand CRM.

A. On-Demand CRM

On-Demand CRM is a low cost customer relationship management solution delivered over the Internet via the cloud. “Customer relationship management (CRM), is a model for managing a company’s interactions with current and future customers, involves using technology to automate and

synchronize sales, customer service and technical support" [8]. These are comparatively cheaper services compared to traditional On-Premise CRM solutions, as they do not require installation at the place of business. There is no maintenance or upgrade costs either, as the cloud provider of these services is responsible for maintaining these functions. The service is viewed as On-Demand, with customers paying only for what they use.

B. Cloud Computing

Cloud computing is considered a new technology trend as well as a disruptive technology¹. There are many definitions for cloud computing. "Cloud Computing provides the facility to access shared resources and common infrastructure, offering services on-demand over the network to perform operations that meet changing business needs" [1]. Hence, customers will be able to access applications and data from a "cloud" anywhere in the world on-demand. It also provides facilities for users to manage their applications on the cloud, which entails the virtualization of resources [9]. However, these definitions are subjective and based on the perspectives they have been derived from. From the point of view of the users or consumers who actually use the services offered through the cloud, cloud computing could be defined in terms of the low cost, fast access, reliable form of service. From the point of view of vendors or suppliers, it is simply a model for providing and renting computer resources in a bid to optimize their use [10]. From a business point of view, it is a new way of reaching out to customers and providing services at a fast and reliable speed. Reference [11] describes SAAS as applications that are custom built as per business needs. They are provided to the end users via clouds that can be public, private or hybrid without the need to manage the platform on which they run or installing them on users' machines. An example of this is Salesforce.com CRM.

C. Cloud CRM Architecture and SAAS

SAAS is the basis of cloud CRM. The striking feature that makes it faster and more flexible when compared to On-Premise CRM is its multitenant architecture in a web-based framework. Multi-tenancy is an architecture framework in which a single instance of a software application serves many customers in isolated instances for each user. Each customer is called a 'tenant'. Tenants may be given the ability to customize some parts of the application, such as the page layout or the user interface, but they cannot change the application code. This is a kind of "one-to-many" model, whereby an application is shared across multiple clients. This framework promotes flexibility and greater speed, as there is no need to provide separate software for each tenant. They all have the same copy of the application, which they can access and customize, but each with a different user login. Hence, instead of installing and maintaining the software, users can access it

via their internet browser, which frees the providers and businesses from complex and time-consuming software and hardware management. Thus, SAAS infrastructure and technology, alongside cloud computing, helps to improve the quality of CRM implementations and address CRM as a Service or "Hosted CRM".

D. Scope And Limitations

The project aims to narrow down the research in order to categorize cloud CRM specifically and specific end user usability attributes that influence the selection decision. The focus is limited to evaluating technological and end user attributes. The architectural factors are not considered extensively because there is already good research available on cloud applications evaluation frameworks, which is highly influenced by cloud architecture and deployment style [7]. In addition, vendor related financial factors were not considered within the scope of this research due to time limitations. As of now, no country limitations feature as part of this research. The remaining sections discuss the related concepts, literature review and research approach and results.

III. RESEARCH APPROACH

Design Science (DS) is a method via which the boundaries of human and organizational capabilities can be extended through the development of innovative artifacts [12]. This paper proposes evaluation criteria for selection of On-Demand CRM systems for SME. SME are goal oriented business entities. One problem in this area is the difference between the goal state and current state of the system. This problem is often solved by designing effective business processes [12]. Hence, a framework in the initial stage of the business process construction, such as pre-selection or selection of a CRM package, plays a major role in enabling effective business processes to achieve these goals. The real essence of the Design Science research method can be realized from the fact that it addresses two fundamental questions: 1) "What utility does the new artifact provide?" and 2) "What does that utility demonstrate?" In addition, the method also requires that the artifact or research adequately maps to the real world; in other words, this approach requires rigor. If it does not then there is no utility as it is not implementable. Thus, if there is no utility then its broad acceptance and future contribution is limited. The reasons highlighted above are the prime motivators for adopting this approach for research in the CRM domain regarding cloud-based software selection, as this research clearly addresses a real world problem and consider a tangible solution, or at least an implementable solution, which has been well evaluated as a part of the research. Reference [12] emphasizes on two very important stages of research. They are Develop/Build and Justify/Evaluate. A link is created between these two stages by focusing on the need to evaluate and justify the propositions so that they can be added to the existing knowledge base. The beauty of this approach is also understood from the fact that it challenges researchers to assess the relevance of the research topic they have chosen and its alignment with business needs within the organizational

¹ Technology or innovation that transforms business and global economy. <http://www.forbes.com/sites/gregsatell/2014/01/05/why-the-cloud-just-might-be-the-most-disruptive-technology-ever/>

context. The methodology starts with a literature review to understand various parameters. A survey in the form of questionnaires is distributed to the CRM audiences including varied mix of respondents from all three perspectives i.e. Customer, provider and user. Quantitative analysis is conducted on the survey data collected. Correlation and Means method is adopted for analysis.

IV. LITERATURE REVIEW

Reference [14] brings to light the strategic perspective that requires concerned stakeholders to take into account the cloud architecture and advancements while assessing the CRM solutions. Hence, any evaluation for On-Demand systems would lie at the intersection of cloud computing characteristics and CRM. Thus, the first task is to identify various stakeholders and their perspectives that will influence this decision. Reference [13] suggests there are two types of stakeholders in the area of CRM applications. One is the “consumers”, which includes the companies that implement the CRM On-Demand solutions, IT investors and decision makers. Second is the “providers”, which includes the vendors that provide the hosted On-Demand CRM solutions in the cloud-based architecture. However, another set of key stakeholder are “end users” of applications, who might be technical staff or non-technical teams such as sales or marketing. These users can no longer be termed as “web users”. It may be more appropriate to call these CRM users “cloud users”, considering the cloud infrastructure through which they access these applications. Thus, evaluation of On-Demand CRM needs to take into account these key sets of stakeholders and users and evaluate it from user perspective, too. Reference [15] propose the key quality factors, SAAS features and metrics for evaluating and assessing these quality attributes as based on IEEE 1061 standard which is the standard software quality evaluation methodology. The factors that they identified are reusability, customizability, data management, scalability, availability and pay per use. Thus, the attributes proposed can be seen as crucial from the cloud architecture point of view. Pay per use is a factor that is a significant economic consideration for a SME. Reference [15] work points out that in case of cloud-based SAAS applications; the software itself is a target of reuse. Reference [15] argues that reusability is no longer limited to the number of modules that can be re-used for different requirements. They associate scalability and availability to the properties of the process and network, rather than simply the system.

Reference [16] maintains that any CRM package is ultimately a software product and the product quality will highly influence any selection decision regarding the software. The product quality evaluation therefore forms an important part of any CRM evaluation. From an architectural point of view, reliability, control of data, integration, high availability and security parameters are of utmost importance to any cloud service evaluation. Although their work and framework can be generalized for any cloud application, some CRM specific evaluation attributes still need to be taken into consideration

when making the selection decision. In the next section, we will consider cloud evaluation attributes in the context of On-Demand CRM.

Many authors suggest other non-technical attributes as being part of CRM evaluation. These are vendor market presence, TCO (total cost of ownership), pricing model or cost, vendors' roll-out time, service training and support, geographic reach and support, change management, management commitment, etc. [17][18][19][10][14]. As previously highlighted, these non-technical aspects have been thoroughly addressed in the literature. Reference [4] points out that the technology factor will be an enabler of entire business process and therefore needs to be constantly studied and addressed according to advancements in technology. Taking into consideration the scope of this research, we focus on the technical factors that affect the CRM On-Demand evaluation process. Vendor related attributes are not considered, as the expertise and dependability of vendors are context dependent and may vary according to geographic region and partner networks. The functional attributes of On-Demand CRM have been widely investigated in the literature and as such, all the functionalities present in On-Premise CRM are mirrored by vendors in On-Demand CRM. Reference [20] suggests that organizations must obtain an informed view of technology and perform technical evaluations to aid in a timely and balanced decision regarding CRM evaluation. The reason for emphasis on the technical variables is that many managers who are typically in-charge of the final selection of the CRM On-Demand application find it difficult to identify the technical variables. Reference [21] argues that organizations and managers have left much of the understanding of the technical IT artifacts to the technology vendors, for example cloud vendors. This view has been reinforced by the large number of whitepapers published by many cloud vendors discussing the technical details of the products and technology and their overall market implications [19][14][22]. However, it can be argued that viewing the assessment parameters through lens of the cloud CRM vendors may not always be the best decision, given the probability of vendors' biases and their intentions to market the product and technology they aim to sell.

Reference [21] points out that there does exist a research gap, particularly in the IS (Information Systems) field, which renders technical variables not part of management decision-making activities. We have therefore excluded the functional and other non-technical evaluation attributes related to process and people from the scope of this research.

V. EVALUATION CRITERIA

A. Evaluation Attributes In Context Of On-Demand CRM

- 1) *Reliability*: The degree to which a software product maintains a specified level of performance under specified conditions is referred to as its reliability [23]. In the context of On-Demand CRM, this reliability is the ability to ensure constant operation

of the system without disruption, i.e., no loss of data, no code reset during execution, etc. It can be argued that there is a strong relationship between availability and reliability; however, reliability focuses in particular on the prevention of loss (of data or execution progress), which is vital in the case of CRM systems.

- 2) *Availability & Robustness*: Availability is the ability of introducing redundancy for services and data so that failures can be transparently masked. With increasing concurrent access as a result of an increased number of CRM users logging in from a variety of devices such as mobile phones, web computers and tablets; availability is particularly achieved through replication of data/services and distributing them across different resources to achieve load-balancing.
- 3) *Scalability*: The ability of the On-Demand CRM application infrastructure to handle a growing number of resources accessing it is referred to as its scalability. On-Demand CRM is a cloud application and cloud systems focus largely on horizontal scalability (instance replication) rather than scalability (changes in the resource structure) [24].
- 4) *Security, Privacy and Compliance*: On-Demand CRM houses customer and critical business data. The security and privacy of data is observed as one of the top challenges of cloud CRM systems [25]. The most obvious concern is lack of control over data and code distribution. It can be argued that an attribute such as multi-tenancy, which is an important enabler of the provision of scalability, can also be seen as a disabler for security and privacy purposes. This is because multi-tenancy allows multiple users to access the same copy of an application in different instances. This could invite untrustworthy entities who may misuse the infrastructure for hacking and DOS (Denial of Service) attacks. In hosted CRM software, all data and applications are hosted in vendor servers in the cloud, and vendors are responsible for its security. Reference [25] argues that this might not go well with some organizations like financial banks and health enterprises, where confidentiality of data is a prime concern. Hence, such organizations may not opt to relinquish control of all their business data to external cloud providers. Mirrored data-centers, vendor backup and disaster recovery mechanisms in service level agreements (SLA), security audits and security attestation such as ISO27001 are seen as solutions for the security purposes by various groups. Jurisdiction of hosted data, legislation models in different countries and compliance of cloud providers with regulations in specific countries are also seen as issues (EU, 2010). The relative importance of this attribute cannot be side-lined, considering the nature of CRM business. Security and privacy concerns are

seen as a never-ending issue in almost all web-based CRM applications. However, separate dedicated teams can be made available to address these issues.

- 5) *Integration*: In an InformationWeek research survey, Reference [26] argues that although integration facilities are provided by cloud CRM vendors, integration with non-SAAS applications remains a top SAAS challenge (62%), as voted by 159 companies who use or are planning to use SAAS. CRM On-Demand is a SAAS product; hence, integration tools provided by cloud applications forms an important evaluation criteria for On-Demand CRM. This is because many CRM systems interact with legacy systems or pool data in and out from legacy systems.
- 6) *Usability*: Reference [23] defines Usability for software products as the capability of the software product to be understood, learned, used and be attractive to the user when used under specified conditions. Usability is an inherent and measurable property of all interactive digital technologies. The users are not simply the end users but also business users or non-technical users such as a sales team, marketing team, etc., for whom the ease of use of a system and its operability will enhance their time for marketing; thus, improving the efficiency of business processes.
- 7) *Customization*: It is defined as the ability to tailor the system according to business needs [23]. This can mean tailoring or customizing the business process functionality or the usability features. This is broadly referred to as configuration in within the context of CRM. Reference [22] argues that the customization ability of the On-Demand model is restricted to a codeless basis as a result of security reasons. Hence, one attribute may sometimes suffer at the cost of making others better. However, with the high level of complexity of business requirements observed in this agile work era, this attribute has gained due importance.
- 8) *Mobility*: Reference [19] defines mobility for CRM applications as the ability of the CRM solution to support a mobile workforce. This includes the breadth of platforms supported and online and offline capabilities according to the functionality addressed. Social media integration is another current market differentiator. Some features include mobile device application and support, iPad and tablet support, cross-browser CRM application support and social media integration (Facebook, Twitter, LinkedIn) with CRM applications. Real-time access to a CRM system through mobile phones (WAP, WML and voice), real-time access to CRM system through PDAs, Palm, Windows and BlackBerry devices etc., are vital for cloud-based CRM evaluation. Most

CRM applications today cannot ignore this attribute, as it truly functions as a market differentiator.

- 9) *Multi-tenancy*: The framework acts as an evaluation criteria, as it is based on cloud architecture. In this way, the same resource is assigned to potentially multiple users at the same time in multiple isolated instances. Information is maintained in separate tables for each user at a database level. However, there are variants of this option available for cloud systems. Discussion of these variants is not within the scope of this research. These are also not applicable to CRM systems, as most instances of application of CRM in the cloud are SAAS based with the multi-tenant variation of web-based architecture applied. Hence, this attribute is part of cloud architecture evaluation. Since most vendors provide CRM applications with the same variant of multi-tenancy, we wish to skip this as an evaluation attribute for On-Demand CRM.
- 10) *Multi-Channel Support*: The solutions ability for supporting a diverse array of communication channels such as email, in person, telephone, web portal interactions and social media can be defined as multi-channel support for CRM. All channels must be tightly integrated with the CRM application so that agents can handle client tickets coming in from all channels (emails, chat, telephone, social networks and web) without switching to any third party applications, i.e., he/she should be able to chat, reply to emails, take/dial calls, etc., all within the same application. Through recent research, Reference [27] concludes that high-quality seamless customer experience is impacted strongly by mobile and online communication channels. However, they warn SMEs not to neglect offline features and to make sure online and offline features are well connected, as the experience of one channel directly influences the customer's willingness to engage via another channel.
- 11) *Reusability*: The ability of the software code to be re-used to serve multiple applications and thereby eliminating reworking and redundancy is referred to as its reusability [23]. Reference [15] argues that in the case of On-Demand CRM, the entire SAAS solution can be viewed as a unit of reuse provisioning services to different companies using the same On-Demand application with customizations as required according to user needs. Hence, this evaluation attribute is not considered particularly important in the case of On-Demand CRM evaluation.
- 12) *Data Management*: As per cloud SAAS infrastructure for CRM applications, many users access the same copy of an application. Many users can customize, add or remove data according to their business needs. Sometimes, in order to achieve scaling, data is distributed flexibly across multiple data sources. At

the same time, the system needs to be aware of data location. The size of data may change at any time with the help of scalability; hence, data management is a crucial evaluation attribute for almost all cloud-based applications. This can also overlap with the security and privacy attribute of evaluation. It is also a crucial concern in the 'age of Big Data'. This aspect is, however, now more important at the vendor side, as they might have different ways of managing data [24]. Hence, we classify this attribute as vendor dependent, too, though it is impacted by the cloud architecture.

- 13) *Analytics*: In this era of Big Data, the volume of data continues to grow on a daily basis. Data also continues to grow in CRM applications. This in turn gives even more importance to analysis tools. Data analysis can yield insight into businesses for enhancing their competitive advantage [19]. This can help non-technical teams such as sales and marketing teams to generate leads and opportunities more efficiently and drive them to completion. These analytic features are current needs and should be part of even basic editions of On-Demand CRM applications. These analytic tools can also be used to conduct historical and comparative trend analyses to gain insight into emerging opportunities and critical issues. Reference [28] claim that these analytic features can help turn information into insights and turn these insights into results. One example is Dashboards within which, users can access a deeper analysis and specific records for diagnosing issues. Predictive analysis and data mining processes embedded in analytics features can help organizations gain additional business intelligence. These features can turn simple and huge volumes of data into Smart Data that enhances CRM productivity.
- 14) *Reporting Tools*: Following analysis of the huge volume of data, the requirement is to make this data available in the form of reports for easy understanding and tracking. This involves features in CRM application that facilitate creation, modification of reports and exporting, formatting, filtering, sorting and applying logic to data contained within the reports.

VI. RELEVANT THEORIES

This research is based on the DS research approach proposed by Reference [12], which provides guidelines for researching in the information systems field to devise innovative artifacts. Cloud-based On-Demand CRM is a new technology trend and can be considered an innovation in itself, due to SAAS delivery mode of service. Reference [29] proposed a theory of Information Systems called the Technology Acceptance Model (TAM), which posits that

perceived usefulness² and perceived ease of use of a system determines an individual’s intention to use the system. Reference [29] argues that new technologies are complex to a set of users who are already used to a particular system and a way of doing tasks. This gives rise to an element of uncertainty in their minds, which has a direct impact on the rate adoption of these technologies. Cloud- based On-Demand CRM is identified as one of the disruptive technologies that have transformed the way of doing business [30]. We therefore logically conclude that the evaluation of On-Demand CRM should also focus on attributes that will enhance the perceived ease of use of the system, which in turn will enhance its usefulness and rate of adoption by SME users. This attribute is identified as usability.

In another theory called “Diffusion of Innovations”, Reference [31] argues that innovation’s³ relative advantage has a direct impact on the rate of adoption of new technologies and systems. In this case, SMEs can be considered the unit of adoption or organizations. Relative advantage can be described as the potential benefit gained by adopting the innovation as opposed to using an existing idea or system [31]. Customization, analytics features and multiple channels of accessibility can be seen as innovative attributes in On-Demand CRM. The relative advantage gained can be considered in terms of increased customer satisfaction, faster cycle times and increased ROI for SME. These innovation attributes directly enhance the rate of adoption of On-Demand CRM by SMEs. Taking into account reference [31] theory, it can be said that the evaluation of On-Demand CRM will have a direct effect on its rate of adoption in SMEs. This is because the main goal of SMEs looking for cloud-based CRM solutions is to make the best use of cloud-based CRM On-Demand systems in order to enhance their business productivity and profitability. Reference [31] points out that any innovation exhibits some characteristics, for example, compatibility, complexity, trialability and observability, and that these characteristics affects its rate of adoption by SMEs. We have already discussed the relative advantage characteristic separately, as it can be measured more in economic terms, as previously discussed. We therefore excluded it from the list of evaluation and discussed the rest of the characteristics. By applying reference [31] theory, we also propose another evaluation attribute, i.e., “Free Product Trials”, which were offered by cloud CRM vendors to SMEs. However, we consider this attribute to be more vendor-specific.

VII. ANALYSIS AND RESULTS

A. Mean

The means of the proposed evaluation criteria indicate the average of the importance on 1-5 Likert scale as seen by the respondents. The criteria sorted according to the means

² The extent to which an individual perceives that using the ‘object’ will increase the performance of the individual [29].

³ An idea, practice or object that is perceived new by an individual or unit of adoption [31].

(importance) are shown in the table I below. According to the respondents, the most important criteria from an architectural perspective are availability, reliability and security as shown in table I below. The most important evaluation attributes from the technical quality perspective are observed as usability and mobility with means above 4.0. Looking at the means below, it can be inferred that none of the attributes proposed as a part of the evaluation framework were seen as unimportant by the respondents.

TABLE I. MEANS OF ON-DEMAND CRM EVALUATION ATTRIBUTES

Evaluation Attribute	Mean	Importance
Reliability	4.23	Very Important
Availability	4.07	Very Important
Scalability	3.97	Important
Security	4.27	Very Important
Usability	4.47	Very Important
Integration	4.00	Important
Customization	4.00	Important
Mobility	4.03	Very Important
Multi-Channel Support	3.83	Important
Analytics	4.00	Important
Reporting Tools	3.97	Important
Free Product Trials	4.19	Very Important

B. Correlation

Pearson’s correlation is used to explore the relationship and strength of the relationship between attributes. On careful interpretation of the quantitative statistical output, it can be said that there was no negative correlation observed among any variables indicating that none of the attributes reverse the importance of other attributes. Hence, all attributes are deemed appropriate for the evaluation of On-Demand CRM. The value of Pearson’s correlation coefficient indicates the strength of the relationships. Reference [32] suggests that for careful understanding of the strength of the relationships, the guidelines that can be applied which group correlations strength as small (r-value=0.10 to 0.29), medium (r-value= 0.30 to 0.49) and strong (0.50 to 1.0) [33]. The correlation results are shown in table II below. Strong positive and moderate positive correlations among the attributes are selectively shown below.

TABLE II. CORRELATION OUTPUTS FOR ON-DEMAND CRM EVALUATION ATTRIBUTES

Attribute	Pearson’s correlation(p)	Strength
Availability		

Reliability	0.730	Strong Positive
Scalability	0.686	
Security	0.652	
Reporting tools	0.652	
Reliability		
Availability	0.730	Strong Positive
Scalability	0.697	
Security	0.737	
Analytics	0.677	
Reporting Tools	0.605	
Scalability		
Availability	0.686	Strong Positive
Reliability	0.697	
Security	0.602	
Security		
Availability	0.652	Strong Positive
Reliability	0.737	
Scalability	0.602	
Reporting Tools	0.789	
Integration		
Customization	0.429	Moderate Positive
Mobility	0.384	
Customization		
Mobility	0.627	Strong Positive
Multi-Channel Support	0.632	
Usability		
Multi-Channel Support	0.646	Strong Positive
Analytics	0.612	
Mobility		
Customization	0.627	Strong Positive
Multi-Channel Support	0.760	
Multi-Channel Support		
Customization	0.632	Strong Positive
Usability	0.646	
Analytics		
Reporting Tools	0.811	Strong Positive
Usability	0.612	
Reporting Tools		
Security	0.789	Strong Positive
Analytics	0.811	
Free Product Trials		
Availability	0.399	Moderate Positive
Reliability	0.577	
Usability	0.595	

C. Results

The results indicate that the proposed evaluation attributes from a technology perspective for On-Demand CRM systems will have a direct impact on the implementation. This is substantiated by the high mean values of above 4.0 for attributes such as usability, security, mobility and availability. The highest mean value of 4.47 for usability indicates that user perspective directly impacts the technology perspective. This is consistent with the AMR Research survey results that indicated 47% of companies reported serious challenges with end user adoption [6]. The mean results and analysis of variances across the SMEs surveyed indicated that none of the attributes proposed were seen as completely inappropriate by the SMEs for evaluation. This substantiates the suitability of the chosen attributes and validates our two main objectives of

identifying and proposing evaluation criteria for cloud CRM for SMEs from a technical quality and user perspective. The two attributes with strongest positive correlations between them were reporting tools and analytics. These attributes are utmost important in this era of Big Data as the businesses in order to enhance their competitive advantage need to capture markets through good analysis of huge data at their disposal [19]. However, in order to gain proper insights from these data and turn them into results and prospective business opportunities, a good reporting feature is inevitable. This is because managers who analyze and transform this data into results need a user friendly interface to transform and understand the data. This all has to be well integrated in the same system. Many participants suggested Free product trials as an evaluation attribute. This was observed as an attractive feature that lets SMEs test and experience the CRM cloud system and its capabilities well before buying them and perform a suitability check. However, in cases where it doesn't fit the company's needs, customization abilities can be considered as a deciding feature. This is illustrated by the fact that customization to suit client needs was considered as an important attribute by all organizations in the survey irrespective of the size of the organizations. The research findings add to a growing body of literature on cloud-based On-Demand CRM and evaluation. This meets one of the guidelines of the Design Science research approach that requires the research conducted should make a good contribution to the existing knowledge base [12].

VIII. CONCLUSION

In this paper, we proposed On-Demand CRM evaluation criteria from a technology and end-user perspective. This can ultimately enable decision makers to make a sound choice of a cloud-based On-Demand CRM solution for their business needs. The initial literature review helped us extract appropriate On-Demand SAAS based CRM evaluation attributes and cloud evaluation attributes. These were then filtered as per the scope of the research. Evaluation criteria was tested, validated and evaluated through a survey and quantitative analysis.

IX. FUTURE WORK

In future, the participants could be interviewed as well as surveyed using a mixed method approach to gain a deeper understanding of the subject. This study can be extended to focus only on large organizations and the results can be compared with the results of the SMEs to gain a better understanding of the evaluation criteria for cloud CRM. In addition to this, the research could also be conducted across the organizations in a specific country. As highlighted earlier, the people and process factors were not considered within the scope of this research. However, CRM as a business strategy is a mix of three factors: people, process and technology. The incorporation of the proposed evaluation criteria with the people and process factors such as functionalities offered by CRM system, cost etc. can provide a more complete and robust framework for evaluation.

ACKNOWLEDGMENT

The author is grateful to the respondents and experts who aided in carrying out research for this work. The author expresses sincere gratitude to Dr. Barnaby Martin and Middlesex School of Science & Technology for their support and guidance.

REFERENCES

- [1] Kulkarni, G., Gambhir, J. & Palwe, R., "Cloud Computing-Software as Service," International Journal of Cloud Computing and Services Science (IJ-CLOSER), vol. 1, no. 1, pp. 11 – 16, 2012.
- [2] Gartner. "Predicts 2013: CRM Goes More Cloud, Becomes an App, Has a New Leader and Changes Name", available online at: <https://www.gartner.com/doc/2264615/predicts--crm-goes-cloud>.
- [3] EU Commission. "EU Enterprise and Industry: What is an SME? ", available online at: <http://ec.europa.eu/enterprise/policies/sme/facts-figures-analysis/sme-definition/> .2003.
- [4] Chen, I. & Popovich, K., "Understanding customer relationship management (CRM) People, process and technology", Business Process Management Journal, vol.9, no. 5, pp. 672-688, 2003.
- [5] Overby, S., "The Truth About On-Demand CRM", CIO Magazine, 2006.
- [6] CRM Landmark, "How big is the CRM software-as-a-service industry? How fast will it grow? When will it cap? ", available online at: www.crmlandmark.com/saasmarket.htm, 2013.
- [7] Reixa, M., Costa, C. and Manuela, A., "Cloud services evaluation framework", In Proceedings of the Workshop on Open Source and Design of Communication, ACM, pp. 61-69, 2012.
- [8] R.Shaw, "Computer Aided Marketing & Selling", Butterworth-Heinemann, Oxford, 1991.
- [9] Gajala, C., "Cloud Computing: A State of Art of the Cloud", International Journal of Computer Trends and Technology, vol. 4, no.1, pp. 35-38, 2013.
- [10] Miguel, R., Carlos, C. & Manuela, A., "Cloud Services Evaluation Framework", In Proceedings of the Workshop on Open Source and Design of Communication, pp.61-69, 2012.
- [11] Hoefler, C. N. & Karagiannis, G., "Taxonomy of cloud computing services", GLOBECOM Workshops (GC Wkshps), IEEE, pp.1345-1350, 2010.
- [12] Hevner, A. R., Jinsoo, P., March, T. S. & Ram, S., "Design Science in Information Systems Research". MIS Quarterly, vol. 28, no.1, pp. 75-105, 2004.
- [13] Thanawin, R. & Veeragandham, M., "CRM: Software as a Service versus On-premise—benefits and drawbacks", 2009.
- [14] Saugatuck Technology, "Toward a Framework for Evaluating Cloud-based CRM", 2011.
- [15] Lee, J. Y., Lee, J. W., Cheun, D. W. & Kim, S. D., "A Quality Model for Evaluating Software-as-a-Service", IEEE Computer Society, pp. 263-266, Washington, DC, 2009.
- [16] Colombo, E. & Francalanci, C., "Selecting CRM packages based on architectural, functional, and cost requirements: Empirical validation of a hierarchical ranking model", Requir. Eng., vol. 9, no. 3, pp. 186-203, 2004.
- [17] Almotairi, M., "CRM Success Factors Taxonomy", European and Mediterranean Conference on Information Systems(EMCIS2008), pp.29-35, 2008.
- [18] Eid, R., "Towards a Successful CRM Implementation in Banks: An Integrated Model", The Service Industries Journal, vol. 27, no.8, pp. 1021-1039, 2007.
- [19] Ovum, "Ovum Decision Matrix: Selecting a CRM Vendor in the Life Sciences Industry", 2012.
- [20] Brown, A. W. & Wallnau, K. C., "A Framework for Systematic Evaluation of Software Technologies", IEEE Software, vol. 13, no.5, pp.39-49, 1996.
- [21] Orlikowski, W. J. & Iacono, S. C., "Research Commentary: Desperately Seeking the "IT" in IT Research- A call to theorizing the IT Artifact", Information Systems Research, vol. 12, no. 2, pp. 121-134, 2001.
- [22] Sage CRM, "On-demand or on-premise CRM: 5 things to consider before making your decision", available online at: https://community.sagecrm.com/user_community/m/whitepapers/3473.aspx, 2010.
- [23] Khosravi, K. & Gueheneuc, Y.G., "A Quality Model for Design Patterns", German Industry Standard, 2004.
- [24] EU, E. C., "The Future of Cloud Computing", 2010.
- [25] Pombriant, D. & Greenberg, P., "The top on-demand CRM and SaaS CRM FAQs", available online at: <http://searchcrm.techtarget.com/feature/The-top-on-demand-CRM-and-SaaS-CRM-FAQs>, 2009.
- [26] Weier, H., "SaaS Integration: Real-World Problems, And How CIOs Are Solving Them", available online at: <http://www.informationweek.co.uk/services/saas/saas-integration-real-world-problems-and/211200952>, 2008.
- [27] IBM, "Delivering a seamless experience across every channel", New York: Econsultancy, 2013.
- [28] IBM, "Business analytics in the cloud", IBM Industries, New York, 2012.
- [29] Davis, F. D., "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology", MIS Quarterly, vol. 13, no. 3, pp. 319-340, 1989.
- [30] Infosys, "Cloud Computing", vol. 7, no. 7, 2009.
- [31] Rogers, E. M., "Diffusion of innovations", 5th ed., Free Press, New York, 2003.
- [32] Pallant, J., "SPSS Survival Manual, A step by step guide to data analysis using IBM SPSS", 5th ed., Maidenhead: Open University Press, 2013.
- [33] Cohen J., "Statistical Power analysis for the behavioural sciences", 2nd ed., Lawrence Erlbaum associates, New York, 1988.

ENSURING SMART GRID DATA SECURITY AT CLOUD DATA CENTRES

E.Chaitanya Krishna

Asst.System Engineer,
Tata Consultancy Servies,Bangalore,India

Dr. K. Venkataramana

Department of Computer Science,
KMM Institute of Post Graduate Studies,Tirupati
Email:ramanakv4@gmail.com

Dr.Sulaiman AlMuhteb

Department of Computer Science,S.V.University,Tirupati

Prof..M.Padmavathamma

Department of Computer Science,S.V.University,Tirupati

Abstract: In the world of evolving technologies the energy like solar or electricity and utilities industry, plays major role includes smart meters and smart grids, which provides companies with exceptional capabilities for forecasting demand, determining customer usage patterns, preventing outages, minimizing the loss and more. Advances in technologies and its usage generates unprecedented data volume, speed and complexity which should be preserved securely for later usage for accurate predictions. Managing the large volume information generated by short-interval reads of smart meter data by various smart devices is a challenge for existing IT resources in storing them and also ensuring the privacy of sensitive customer meter data is also a major issue in smart meter deployments. Security should be provided for data which is stored at data centers at cloud and also at local energy distributor centers at two tiers. In this paper we focus on security regarding storage of big data at data centres as well as at local distribution such as databases. So we propose MDET (Multiple Data Encryption Technique) which allows encryption of each record two times at storage centres by using Generator based encryption technique. In this technique the data in the database is encrypted twice so that the data record should be decrypted once at data centre as well as at the local distribution centre by private keys at two levels so that privacy of consumer data is not lost at either data storage centre at cloud or at local distribution centre.

Keywords---Smart grid, Goals of smart grid, Functions of smart grid, Securing Grid data, Big Data centres, Encryption and Decryption.

I. INTRODUCTION

A **Smart Grid** as a digitally enabled electrical grid that combines modern IT technology with Electrical system that gathers, distributes and acts on information about the behavior of all participants (suppliers and consumers) in order to improve the efficiency, importance, reliability, economics, and sustainability of electricity services . The deployment of advanced metering infrastructure (AMI) and intelligent supervisory control and data acquisition (SCADA) systems is essentially all about improving the amount and quality of data that utilities have on supply and distribution. Data is the fundamental currency of the smart grid. A clear understanding of how this data is generated, what it consists of and the benefits it can be used to deliver is critical to realizing the fullest possible returns from

smart grid investments. Over the past 50 years, electricity networks have not kept pace with modern

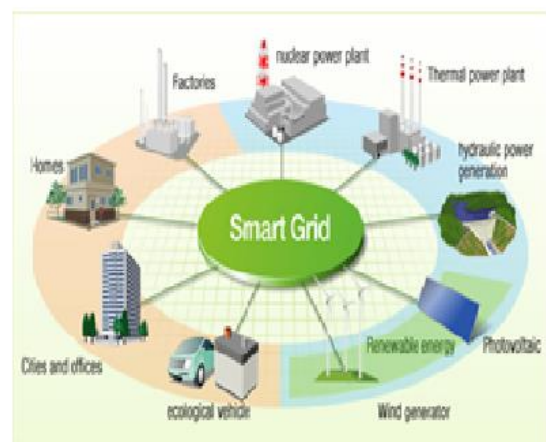


Fig-1 SMART GRID

challenges, such as: Data storage, Security threats, from either energy suppliers or cyber-attack, national goals to employ alternative power generation sources whose intermittent supply makes maintaining stable power significantly more complex. In general, data management design in any context should optimize outcomes in two ways. First, it should extract clean, consistent and well understood information that drives targeted benefits for the business. And second—having identified those benefits—it should minimize the costs of infrastructure needed to obtain and process the data necessary to deliver these benefits. Before inception of Smart grids the data relating to distribution of electricity collection is very rudimentary like consumption data on customer premises one data point a month per customer. But Smart grids have changed things. The deployment of advanced metering infrastructure (AMI) and intelligent supervisory control and data acquisition (SCADA) systems is essentially all about improving the amount and quality of data that utilities have on supply and distribution. The advent of AMI has increased the level of data collection dramatically from megabytes to Exabyte which requires Data centres and cloud infrastructure. Modern Smart grid data should be maintained at data centres at cheap cost, also should be guarded against unauthorized access at service provider level as well as at data centre level [8].

A smart grid is a large-scale system that extends from a power generation facility to each and every power consuming device such as home appliance, computer, and phone. This large-scale nature has increased the possibilities of remote operation of power management and distribution system. With energy being a premium resource, ensuring security against theft, abuse, and malicious activities in a smart grid is of prime concern.

The challenges of ensuring cybersecurity in a smart grid are diverse in nature due to the diversity of the components and the contexts where smart grids are deployed. Deploying a smart grid without strong and diligent security measures can allow advanced cyber-attacks to remain undetected, which can eventually compromise the entire system [9]. Inadequate security measures can also compromise the stability of the grid by exposing it to, for example, utility fraud, loss of confidential user information and energy-consumption data [10].

The cyber security objectives can be classified into the following three categories [9, 11]. (i) Integrity. Protecting against the unauthorized modification or destruction of information. Unauthorized information access opens the door for mishandling of information, leading to mismanagement or misuse of power. (ii) Confidentiality. Protecting privacy and proprietary

information by authorized restrictions on information access and disclosure. (iii) Availability. Ensuring timely and reliable access to information and services. Availability can be compromised by disruption of access to information which undermines the power delivery. To ensure above, for securing smart grid data we have proposed an MDET algorithm for providing privacy and secrecy for data at data centres storage.

II. ARCHITECTURE OF SMART GRID

An electrical grid is not a single entity but an aggregate of multiple networks and multiple power generation companies with multiple operators employing varying levels of communication and coordination, most of which is manually controlled. Smart grids as show in Fig-1 increase the connectivity, automation and coordination between these suppliers, consumers and networks that perform either long distance transmission or local distribution tasks. This paradigm is changing as businesses and homes begin generating more wind and solar electricity, enabling them to sell surplus energy back to their utilities. Modernization is necessary for energy consumption efficiency, real time management of power flows and to provide the bi-directional metering needed to compensate local producers of power. Although transmission networks are already controlled in real time, many in the US and European countries are antiquated by world standards, and unable to handle modern challenges such as those posed by the intermittent nature of alternative electricity generation, or continental scale bulk energy transmission. Smart Grids are profitable for industrial sector in various like aluminum processing, cement manufacturing, food processing etc. which should function in accordance with by greenhouse emission standards. [4] Smart Grid uses computer communication networks which requires security [5].

III. GOALS OF THE SMART GRID

Latency of the data flow is a major concern, with some early smart meter architectures allowing actually as long as 24 hours delay in receiving the data, preventing any possible reaction by either supplying or demanding devices.

A. Smart Energy Demand

Smart energy demand describes the energy user component of the smart grid. It goes beyond and means much more than even energy efficiency and demand response combined. Smart energy demand is what delivers the majority of smart meter and smart grid benefits. Smart energy demand is a broad concept. It includes any energy-user actions to: a) Enhancement of

reliability b) reduce peak demand, c) shift usage to off-peak hours d) lower total energy consumption e) actively manage electric vehicle charging f) actively manage other usage to respond to solar wind and other renewable resources g) buy more efficient appliances and equipment over time based on a better understanding of how energy is used by each appliance or item of equipment. All of these actions minimize adverse impacts on electricity grids and maximize utility and, as a result, consumer savings.

IV. FUNCTIONS OF SMART GRID

Before examining particular technologies, a proposal can be understood in terms of what it is being required to do. The governments and utilities funding development of grid modernization have defined the functions required for smart grids. Smart Grid Technology will have following functionalities [1] such as

A. Self-healing

Using real-time information from embedded sensors and automated controls to anticipate, detect, and respond to system problems, a smart grid can automatically avoid or mitigate power outages, power quality problems, and service disruptions. Technology such as Fault Detection Isolation and Restoration can be used in conjunction with protective relays to automatically detect and isolate a fault, and then restore power to as many customers as possible. This will greatly improve the reliability of the electrical distribution network

B. Consumer Participation

A smart grid is a means for consumers to change their behavior around variable electric rates or participate in pricing programs designed to ensure reliable electrical service during high-demand conditions. Historically, the intelligence of the grid in North America has been demonstrated by the utilities operating it in the spirit of public service and shared responsibility, ensuring constant availability of electricity at a constant price, day in and day out, in the face of any and all hazards and changing conditions. A smart grid incorporates consumer equipment and behavior in grid design, operation, and communication.

C. Resist Attack

Smart grid technologies better identify and respond to man-made or natural disruptions. Real-time information

enables grid operators to isolate affected areas and redirect power flows around damaged facilities.

D. High Quality Power and Generation options

As smart grids continue to support traditional power loads they also seamlessly interconnect fuel cells, renewables, micro turbines, and other distributed generation technologies at local and regional levels. Integration of small-scale, localized, or on-site power generation allows residential, commercial, and industrial customers to self-generate and sell excess power to the grid with minimal technical or regulatory barriers.

E. Enable Electricity Market

Significant increases in bulk transmission capacity will require construction of new transmission lines before improvements in transmission grid management proposed by smart grids can make a difference. Such improvements are aimed at creating an open marketplace where alternative energy sources from geographically distant locations can easily be sold to customers wherever they are located.

F. Optimize Assets

A smart grid can optimize capital assets while minimizing operations and maintenance costs. Optimized power flows reduce waste and maximize use of lowest-cost generation resources. Harmonizing local distribution with inter-regional energy flows and transmission traffic improves use of existing grid assets and reduces grid congestion and bottlenecks, which can ultimately produce consumer savings. Smart Grid technologies will enable power systems to operate with larger amounts of such energy resources since they enable both the suppliers and consumers to compensate for such intermittency.

V. SMART GRID DATA MANAGEMENT

In terms of the flow of smart grid data, we have identified five architectural stages that can be used to guide the design of the data management structure. As Figure 2 illustrates, data is initially generated by network devices such as meters and sensors, before being transported for storage and processing by various applications—the persistence phase. Then it is transformed into actionable operations-oriented information for network and technical analysis, requiring new visualization capabilities. Finally, the resulting analytics applicable for the non-real time operational consumption are integrated at the enterprise level to drive strategic decision making.

There are five separate classes of smart grid data, each with its own unique characteristics.

1. *Operational data*—represents the electrical behavior of the grid. It includes data such as voltage and current phasors, real and reactive power flows, demand response capacity, distributed energy capacity and power flows, and forecasts for any of these data items.

2. *Non-operational data*—represents the condition, health and behavior of assets. It includes master data, data on power quality and reliability, asset stressors, utilization, and telemetry from instruments not directly associated with grid power delivery.

3. *Meter usage data*—Includes data on total power usage and demand values such as average, peak and time of day. It does not include data items such as voltages, power flows, power factor or power quality data, which are sourced at meters but fall into other data classes.

4. *Event message data*—consists of asynchronous event messages from smart grid devices. It includes meter voltage loss/restoration messages, fault detection event messages and event outputs from various technical analytics. As this data is triggered by events, it tends to come in big bursts.

5. *Metadata*—is the overarching data needed to organize and interpret all the other data classes. It includes data on grid connectivity, network addresses, point lists, calibration constants, normalizing factors, element naming and network parameters and protocols. Given this scope, managing metadata for a smart grid is a highly challenging task. While the first three of these classes are relatively familiar to utilities, the last two have been less prominent to date—and are likely to present more problems as utilities adapt to the smart grid world

In our view, there are two prerequisites for overcoming the challenges of the smart grid data deluge. One is ensuring that the five data classes we previously highlighted are reflected in the data integration architecture. The other prerequisite is the effective use of the right analytics to turn the mass of data into usable information and business intelligence.

If designed properly, the data architecture will provide the capabilities utilities will need to deal with future change and evolution in their smart grids and business environment. To do this, the architecture will need to include more than just data stores, but also elements such as master data management, services and integration buses to effectively share data and information.

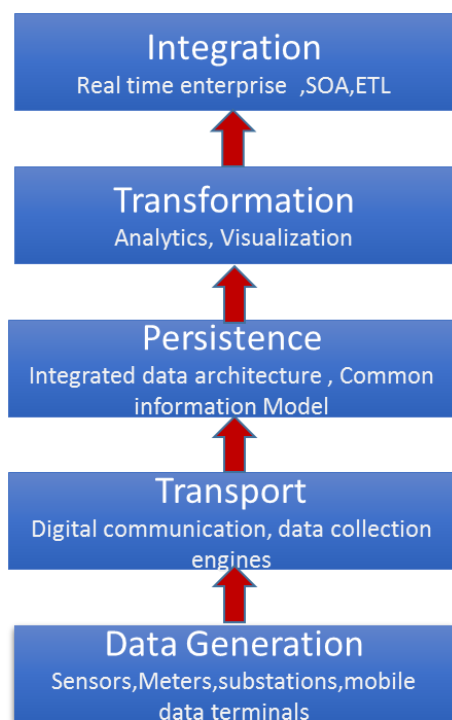


Figure-2 Five architectural stages of smart grid data management.

VI. VULNERABILITIES IN SMART GRID DATA MANAGEMENT

Smart grid network which not only provides improved capabilities to the conventional power network making it more complex in terms of generating huge volume of data which leads to vulnerable to different types of attacks. These vulnerabilities might allow attackers to access the network, break the confidentiality and integrity of the transmitted data, and make the service unavailable [2][3]. The following are the vulnerabilities to be considered:

- a) Network security of distributed systems across meters, substations, poles and In-home devices including authentication, detection, and monitoring
- b) Identity & access management for managing customer information
- c) Messaging and application security communications including data network communications, and transactions.
- d) Security policy management and implementing web services security standards
- e) Customer security: Smart meters autonomously collect massive amounts of data and transport it to the utility company, consumer, and service providers. This data includes private consumer information that might

be used to infer consumer's activities, devices being used, and times when the home is vacant.

f) Greater number of intelligent devices: A smart grid has several intelligent devices that are involved in managing both the electricity supply and network demand. These intelligent devices may act as attack entry points into the network. Moreover, the massiveness of the smart grid network (100 to 1000 times larger than the internet) makes network monitoring and management extremely difficult.

g) Physical security: Unlike the traditional power system, smart grid network includes many components and most of them are out of the utility's premises. This fact increases the number of insecure physical locations and makes them vulnerable to physical access.

h) Implicit trust between traditional power devices: Device-to-device communication in control systems is vulnerable to data spoofing where the state of one device affects the actions of another. For instance, a device sending a false state makes other devices behave in an unwanted way.

i) Different Team's backgrounds: Inefficient and unorganized communication between teams might cause a lot of bad decisions leading to much vulnerability.

j) Using Internet Protocol (IP) and commercial off-the-shelf hardware and software: Using IP standards in smart grids offer a big advantage as it provides compatibility between the various components. However, devices using IP are inherently vulnerable to many IP-based network attacks such as IP spoofing, Tear Drop, Denial of Service, and others [7].

VII. PROPOSED SECURITY MODEL TO SMART GRID DATA

Encryption is mechanism is used to address security or privacy concerns whether is it is a small or big data, or when data in data centres are shared between different types of consumers. The main aim is to secure every clients data so that it will remain inaccessible to unauthorized parties even if they come into possession of it. By above study we have proposed this model shown in figure-3 known as Smart Grid Secure Data Management Model [8].

In this proposed model we enhance the security of Smart Grid Data (SGD) stored data in cloud data centre by providing multi-level layered security solution for data at storage and at the client level. Cloud Service Provider(CSP) stores SGD stored at Data centre is monitored by Security Provider Service (SPS) which

generates keys to ensure security by encrypting data stored at Data centre as well as data stored at Distributer site. In SPS module we have used MDET security algorithm which provides multiple keys to ensure security to customer data.

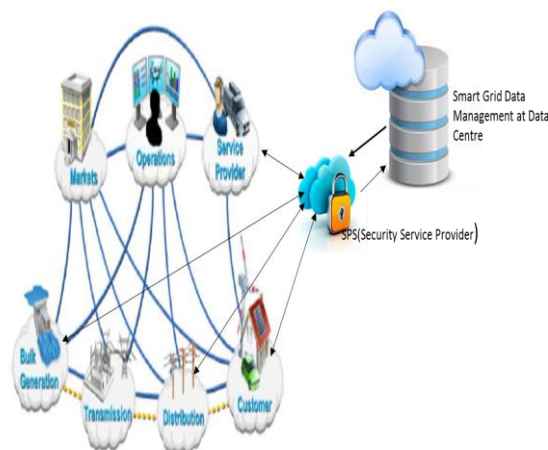


Fig: 3 Smart Grid Data at Data centre Secured by SPS service

The Proposed model mainly contains two Services controlled by CSP

a) *Registration Service* used for registration of Distribution Company Client/Customer which is using Smart Grid Services to store data in data centres in Cloud.

b) *Security Provider Service (SPS)* which authenticates client request and stores clients data securely which will be secured from attacks in networks. In this SPS we have used various MDET which uses cryptographic techniques for secure storage and transfer of data in communication networks.

VIII. MULTIPLE DATA ENCRYPTION TECHNIQUE (MDET)

In this environment of Data Centres at Cloud terabytes of data generated by Smart Grid applications are stored which raises various concerns of security are raised. So to ensure security for data of each client, in this paper we are proposing Multiple Data Encryption Technique(MDET) in which each record in stored is encrypted two times by Security Provider Service (SPS) first by Data centres's public key and by Clients, and decrypted only by Client. Since the record 'R' to be stored at data centre is encrypted by twice, Smart Grid Data SGD cannot be revealed. In this way the proposed

MDET technique is secure by not revealing the record to other client or Intruders at Data centres in Cloud.

In this technique, we assume $C_1, C_2, C_3, \dots, C_n$ are clients of Smart Grid Company who stores data in Data centres in cloud C. Let Cloud service provider (CSP) and Security Service Provider provides sharing of database securely to multiple clients In this algorithm the data is stored in encrypted format for security purpose, as the database is in cloud, to avoid unauthorized users to access data. The algorithm is given below

A. MDET ALGORITHM

1. SPS generates a large Prime T_p from credentials of Client user of Smart Grid Data stored in data centre in Cloud Service Provider.
2. SPS computes $N=2*T_p$
3. SPS generates Cyclic group Z_N^* of order $\phi(N)$ (Euler Totient function)
4. A subgroup $Z_{\phi(N)}^*$ subset of Z_N^* of order $\phi(\phi(N))$ is generated by SPS with generator $g \in Z_n^*$
5. SPS picks randomly picks up two private keys T_q and $T_r \in Z_N^*$ $T_q \equiv g^{k1} \pmod N$ and $T_r \equiv g^{k2} \pmod N$ where $k1, k2 \in Z_{\phi(N)}^*$ where g is generator for Z_N^*
6. SPS computes $N = T_q * T_r$ for Smart Grid Client C_i
7. SPS chooses 'e' for C_i such that $\gcd(e, \phi(N))=1$
8. SPS determines 'd' for C_i such that $ed \equiv 1 \pmod{\phi(N)}$
9. SPS computes $CP_r = e.rs_t$ such that $e.rs_t \equiv 1 \pmod{\phi(N)}$ and $CP_b = d.rs_d$ such that $d.rs_d \equiv 1 \pmod{\phi(N)}$ where CP_r : Private key $\langle CP_r, d, e \rangle$, CP_b : Client public key, Public key $\langle N, CP_b \rangle$
- 10 SPS encrypts the data of each Client C_i stored in data centre with First level key and obtains Clients encrypted record (CER) $CER = R^e \pmod n$
11. SPS stores CER in C_i after encrypting CER another time to obtain Encrypted Smart Grid Data Record $ESGDR = CER^{CP_b} \pmod n$
12. When Clients data is used/requested at Distributor site other than data Centre SPS sends Security Key to Distributor Location separately.
14. When Client C_i requests data from Cloud CSP sends encrypted record ESGDR to Distributor Site.

15. After receiving Client at distributor site C_i computes Smart Grid Data $SGD = ESGDR^{rst} \pmod N$ to obtain original Record.

IX. CONCLUSION

Smart Grid is required that combines Information Technology (IT) with renewable energy to significantly improve how electricity is generated, delivered, and consumed. Smart grids generates huge volumes of data which are stored in data centres in cloud which should be stored securely to avoid modification of data by unauthorized users which may lead to collapse of grid. So the Smart Grid Data (SGD) in data centres should be secured by cloud with an efficient cryptographic technique which we have proposed as Multi Data Encryption Technique (MDET). MDET technique is used by Security Provider Service which encrypts data twice for security. Smart grid data management will enable the information collected through smart grids will not only empower customers to manage their electricity consumption but will enable electricity system operators to better understand and meet users' needs. In this paper we have taken the issue of security to Smart Grid data and given a Model containing SPS which helps CSP to ensure security and in our future works we give practical results to above model.

REFERENCES

- [1] Sinha, A.; Neogi, S.; Lahiri, R.N.; Chowdhury, S.; Chowdhury, S.P.; Chakraborty, N, "Smart grid initiative for power distribution utility in India", Power and Energy Society General Meeting, IEEE, (2011).
- [2] Pearson I. Smart grid cyber security for Europe. Energy Policy, 2011; 39(9):5211-5218.
- [3] Clements S and Kirkham H. Cyber-security considerations for the smart grid. In: Proc of the IEEE Power and Energy Society General Meeting, 2010:1-5.
- [4] Tariq Samada, Sila Kiliccote, "Smart grid technologies and applications for the industrial sector", Journal of Computers and Chemical Engineering, Elsevier, 2012
- [5] Wenye Wang, Yi Xu, Mohit Khanna, "A survey on the communication architectures in smart grid", Journal of computer networks, Elsevier, pp 3604-3629, 2011
- [6] Fadi Aloula*, A. R. Al-Alia, Rami Al-Dalkya, Mamoun Al-Mardinia, "Smart Grid Security: Threats, Vulnerabilities and Solutions, International Journal of Smart Grid and Clean Energy vol. 1, no. 1, September 2012,
- [7]. K.Venkataramana, Prof.M.Padmavathamma, Multi-Tenant Data Storage Security In Cloud Using Data Partition Encryption Technique, International Journal of Scientific & Engineering Research, Volume 4, Issue 7, July-2013, ISSN 2229-5518
- [8]. Jason Deign, Carlos Márquez Salazar, Data Management and Analytics For Utilities, Smart Grid Update, www.smartgridupdate.com

- [9]. E. Hayden, “*There is No SMART in Smart Grid without secure and reliable communications*,” Tech. Rep., Verizon, http://www.verizonenterprise.com/resources/whitepapers/wp_no-smart-in-smart-grid-without-secure-comms_en_xg.pdf.
- [10]. X. Fan and G. Gong, “*Security challenges in smart-grid metering and control systems*,” Technology Innovation Management Review. In press.
- [11]. *Guidelines for Smart Grid Cyber Security*, The Smart Grid Interoperability Panel: Cyber Security Working Group, Gaithersburg, Md, USA, 2010.

Consistency tradeoffs on distributed multi-datacentric systems

SCOLCH theorems

Balla W. Diack and Samba Ndiaye
Dept. of Mathematic and Informatics
Cheikh Anta Diop University of Dakar
Dakar, Senegal 18522
balla.diack@ucad.edu.sn
samba.ndiaye@ucad.edu.sn

Yahya Slimani
Department of Computer Sciences
ISAMM, University of Manuba
2060 Manuba, Tunisia

Abstract—Distributed systems in the cloud computing context spread data across geographically remote datacenters to ensure always availability, scalability, and a best reactivity. Choosing latter properties in these systems leads to consistency issues (version conflicts, obsolete data, etc.); besides, most analytical solutions suggested for these issues are incomplete. SCOLCH proposes tradeoffs to achieve the required properties for service level agreement in cloud computing.

Keywords— multi-datacentric systems; consistency; convergence; latency; availability

I. INTRODUCTION

Data explosion on datawarehouses referred to as BigData, has completely shaken the modern distributed systems and has led to the cloud computing which was impulsed since the mid-2000s by Amazon, Google, Salesforce.com, etc. The most of cloud service providers have set new levels of consistency in their distributed databases (Dynamo [15], PNUTS [13], BigTable [14], Cassandra [19]) to ensure better performance and to keep their databases always available. These actors claimed that the eventual consistency [29] should not be overcome by distributed systems in the cloud. This statement is based on the CAP theorem (Consistency, Availability, Partition tolerance), also known as Brewer's theorem [10]. Nonetheless, number of researchers have criticized this theorem and have showed its limitations [27, 28].

In section 2 of this paper, we give mathematical formalizations of basic concepts in multi-datacentric systems. In section 3, we give new theorems associated with tradeoffs between strong consistency, convergence, high availability, low latency and causal consistency. In section 4, we remind the related works on these issues. In the last section, we conclude and we give some opened issues.

II. WIDELY DISTRIBUTED SYSTEMS ON THE CLOUD

In this section, we outline basic concepts which are often used but rarely explained by the authors and which are subject

to a lot of confusions. Cloud computing consists essentially of a set of datacenters (thousands of servers per datacenter) and services provided to ubiquitous clients across the internet. A datacenter allows to house computer systems and their associated components.

A. Basic concepts

1) *Unreliable and asynchronous systems*: A system is said to be unreliable whether the messages between nodes may be reordered, dropped, or delayed for an arbitrary but finite duration [23]. A distributed system is asynchronous if its logical local clocks run at different speeds, i.e two operations that are performed simultaneously at two different nodes may appear to be executed at different logical time.

2) *Safety and liveness*: Safety in distributed systems means that some bad thing doesn't happen during execution; liveness means that a good thing happens eventually [2, 20].

3) *Operation*: An operation u is either a read operation or a write operation; it is characterized by two timestamps: the beginning of its performing on a node $start(u)$, and the deliverance of a response which indicate the termination of the operation $resp(u)$. An operation belongs to an execution (process), deals with an item on a node in a datacenter.

Notations: We use the following notations which are equivalent: $(op_i(O,x))_{do/N_j}$, $((op_i(O,x))_{do})$, $((op_i(O,x))_{N_j})$, $((op_i(x))_{do})$, $((op_i(x))_{N_j})$, $(op_i(x))$

Whether $op_i(O; x)_{d=N_j}$ is an operation performed on the node $N_j \in d$ a datacenter in the system. O is the item involved in op and x is the value modified or read in O . These depend on whether you point out the datacenter and the node where the operation is performed.

4) *Execution*: An execution (process) is a series of operations performed by a user.

5) *Availability*: It's a liveness property that means that all operations issued to the datacenter complete successfully. No operation can block indefinitely. The high availability means that any operation to a relevant node should result a response.

B. Happen-before relation:

We recall the classical happen-before relation (HBR noted \rightarrow) with adjustments. HBR is defined on a graph (G) which vertices reflect all operations performed by a relevant node in the system. HBR satisfies the following assertions:

- If a and b are two operations on a same execution, then $a \rightarrow b$ if there's an oriented edge from v_a to v_b (with v_a and v_b the respective vertices corresponding to a and b).
- If w is an update and r is a read that returns the value written by w , then $w \rightarrow r$.
- Let $a, b,$ and c are three operations in an execution, if $a \rightarrow b$ and $b \rightarrow c$, then $a \rightarrow c$.
- Two operations u and v are concurrent if $u \rightarrow v$ and $v \rightarrow u$ are not verified.

C. Levels of consistency

1) *Causal Consistency*: For any execution: $U = \{u_1 \dots u_k\}_{k \geq 1}$, U is causally consistent if and only if:

- There is a serial order of the operations of U at each node, i.e., HBR is verified on the operations in U .
- Any read operation r in U at a node N_i on an item o returns the latest concurrent write at this node on o . $r(o): x = x_o$ such as $w_k(o, x_o) \Rightarrow \forall j \neq k w_j(o, x) \rightarrow w_k$.

Causal consistency means that an operation $op_i(o, x_o)_{dk/N_j}$ completes if and only if: $\forall w_i$ such as $w_i(x_o)_{dk/N_j} \rightarrow op_i(x_o)_{dk/N_j}$ then w_i is completed.

2) *Linearizability or strong consistency*: An execution U is said to be linearizable if its operations appear to take effect across the entire system at a single instance in time between the invocation and the completion (delivrance of response) of the operation.

$$\forall w_i, (w_i(o, x_o)) \in U \Rightarrow (r_i(o):x_o)_{dh} \forall d_h \in D$$

D is the set of datacenters $\{d_1 \dots d_m\}_{m \geq 1}$; w_i the write number i of U ; d_h is a random datacenter; x_o is the latest updated value of item o returned w_i .

3) *The window of inconsistency* is the duration in which an item is not up to date at a node.

D. Latency and convergence:

1) *Latency*: It is the delay between a request starts and its completion; particularly, the low latency is the latency which does not exceed few tens of milliseconds.

2) *Convergence*: A system is strongly convergent if any set of relevant and connected nodes that have received, performed and propagated the same updates will have equivalent state, i.e., all the reads on these nodes will return the same result.

$$\sum_{i=0}^m w_i = \sum_{j=0}^m w_j \Rightarrow \forall r, (r(o,x))_{N_i} = r(o,x)_{N_j}$$

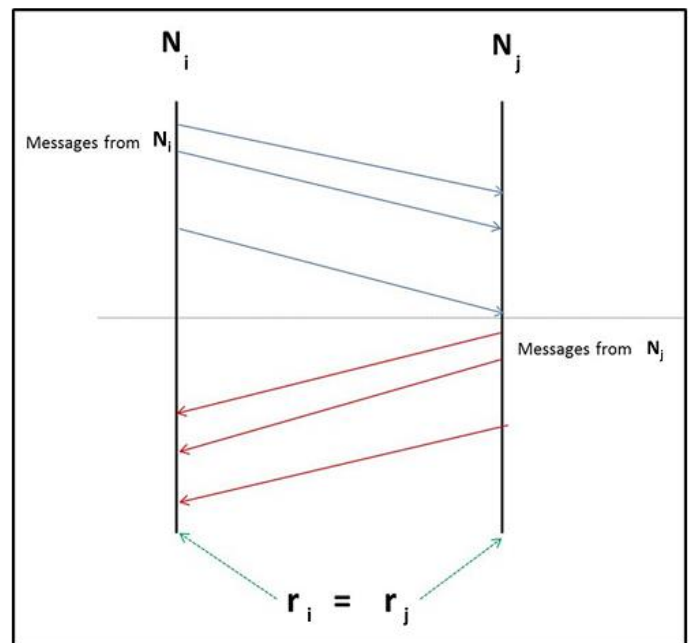


Fig. 1. Illustration of the one way convergence

In this section, we've outlined important concepts, what should be used in the next section to prove our choices.

III. NEW TRADEOFFS ON MULTIDATACENTRIC SYSTEMS

In this section, we are proving the incompatibility between strong consistency in a side and high availability, low latency and convergence in another side. Afterwards, we prove that we can guarantee the latter three properties if the causal consistency is ensured.

A. Strong consistency in multi-datacentric systems

1) *Proposition 1*: "Any multi-datacentric system is unreliable and asynchronous."

The unreliability is intrinsic to widely extended networks as reported by Peter Deutsch [16].

Synchronization means that all nodes in the system have the same clocks (a clock is a function in each node which returns a real number for any operation performed on that node). Ensuring the synchronization requires an enormous overhead, furthermore, in actual multi-datacentric systems; synchronizing clocks on the nodes is not so relevant if we want to guarantee always availability..

2) *Theorem 1*: “Any multi-datacentric system, which guarantees strong consistency, cannot be always available.”

We assume an asynchronous and unreliable distributed system with n servers $(S_1 \dots S_n)_{n \geq 2}$ allocated to m datacenters $(d_1 \dots d_m)$ and we assume that the strong consistency is guaranteed.

$$d_i = \{S_1 \dots S_i\} \dots d_m = \{S_{j+1} \dots S_n\} \quad i \leq j$$

A break on the network between two datacenters $d_p = \{S_1 \dots S_{i+k}\}$ and $d_q = \{S_j \dots S_{j+k}\}$ partitions their set of servers. All messages between the two datacenters are lost or delayed until the recovery of the partitions. We suppose by contradiction that the system provide high availability.

Let u_i be an update on two datacenters (d_p and d_q); $u_i(o,x)_{d_p}$ is performed on a server in d_p and is propagated through the others replicas of o . The high availability implies that the update is committed at all the replicas. Knowing that d_q is unreachable; any read $r_i(o)_{d_q}$ to a replicas in d_q during the inconsistency window will necessarily return a wrong response $x_q \neq x$; which violate the strong consistency. Therefore we came across a contradiction. Consequently, the system will not be always-available if it is strongly consistent.

B. Latency tradeoffs

Although some works had highlighted the latency tradeoff on distributed systems [6,22,23], neither of them has mathematically proven the relationship between the latency and consistency level. In this section we try to prove that.

1) *Theorem 2*: “Any multi-datacentric system that ensures strong consistency will see its latency increasing dramatically.”

We suppose a multi-datacentric system (\dot{S}) , which is strongly consistent. For any execution $U \in \dot{S}$, and for any update $u_i(x) \in U$ on two nodes belonging to any pair of datacenters $(d_j, d_k) \in \dot{S}$, $u_i(x)_{d_j} = u_i(x)_{d_k}$. The latency of $u_i(x)$ is:

$$L_G = \alpha \times (\lambda/c) + L_1 \quad (1)$$

λ : the distance between d_j and d_k

c : the speed of the light, $c \approx 3^8$

L_1 : latency at the datacenter hosting the receiver node.

α : see lemma 1

λ/c is bounded and rarely exceeds 0.1 therefore latency depends essentially on α factor.

2) *Lemma 1*: In unreliable and asynchronous distributed systems, the more the system is spread, the more α is increasing.

Linearizability implies that item replicated to multiple geographically different sites must be up to date at any moment.

$$u(x)_{dk} = x_0 \quad \forall (d_k)_{k \geq 1} \in \dot{S} \quad (2)$$

Where x_0 is the latest value up to date of x

Linearizability requires a lot of message sending between datacenters of \dot{S} , hence we'll have:

$$\alpha = \sigma \times MSG_OP \times IMPL_DTC \times RF \quad (3)$$

MSG_OP: average number of messages per operation

IMPL_DTC: number of datacenters involved in U

RF: A redundancy factor defined to offset the lost messages

σ : a likelihood which increases with the distance between nodes, the number of routers, the number of messages.

It's clear that α increases seriously whether the number of involved datacenters grows (*IMPL_DTC* and σ are increasing).

C. Best level of consistency in multi-datacentric systems

Most cloud providers grant just the eventual consistency and claim that it should not be overtaken [15]. But some recent works proved that eventual consistency should be overcome [7,23]. Here we are showing that the causal consistency which is better than eventual consistency can be guaranteed in highly available systems.

1) *Lemma 2*:

“Any asynchronous and unreliable distributed system can ensure high availability.”

When we have an asynchronous system, termination of operations is not affected because there's different timestamps for each node. Therefore, the availability can be guaranteed since operations should not be blocked. Whether a system is unreliable, messages may be dropped, reordered, or delayed in an undefined but finite duration. This may affect the correctness of the results (safety) but not their outcome (liveness). Furthermore, dropping messages does not preclude the eventual completion of the queries accordingly any operation should terminate. From the above remarks, we can conclude that the high availability can be implemented in such unreliable and asynchronous systems.

2) *Proposition 2*: “Any system which guarantees the causal consistency can accept concurrent writes.”

It follows that for any execution U and G its HBR graph, all the write operations $(w_a, w_b) \in U$ such as $w_a \rightarrow w_b$ and $w_b \rightarrow w_a$ are not verified in G , can be concurrent.

3) *Theorem 3*: “Any asynchronous and unreliable distributed system, which guarantees high availability, can ensure at most causal consistency.”

We suppose a highly available system which ensures a consistency stronger than the causal consistency. For an execution in such systems $U=\{u_1...u_k\}_{k \geq 1}$ and G its associated HBR graph; we proceed as follows:

We construct another execution $U'=\{u_1...u_n\}_{n > k}$ from U by joining it with $R=\{u_k...u_n\}$ a set of read operations which control the implementation of U . We add the vertices corresponding to these reads to G and we obtain a new graph G_I ; then we construct edges on G_I in the following way: for each read in R we build an edge between it and all the non-local write operations in which it depends in G_I . The idea is to come establish a contradiction by processing a large number of operations; either we'll have concurrent write operations necessarily (no operation blocks due to the high availability). This means that:

- $\exists(w_i, w_j) \in (U')$, such as w_i and w_j are concurrent so neither $w_i \rightarrow w_j$ nor $w_j \rightarrow w_i$ should be verified in G_I . Therefore U' does not exceed the causal consistency which tolerate the concurrency between any pair of write operations which are not dependent.
- Finally based upon proposition 2, we can assert that strong consistency cannot be achieved in multi-datacentric systems.

D. Latency, convergence and high availability on multi-datacentric systems

1) *Corollary 1 of lemma 2*: “Any asynchronous and unreliable system which provides causal consistency can achieve a high availability.”

2) *Lemma 3*: “Any multi-datacentric system which achieves causal consistency can guarantee strong convergence.”

- Let $U=u_1, ..., u_k$ be an execution in such system and (w) an update of U on two replicas (R_1, R_2). Write operations are performed if for any operations (op_i) such as: $op_i \rightarrow w$, then op_i is performed before w at each replica. The idea is to commit entirely updates involving items on R_1 and R_2 ($u_i \in U$). And thereafter R_1 exchange messages involving all the updates of U with R_2 .
- Lemma 2 and theorem 3 allows us to assert that finally any read on the items updated on R_1 and R_2 will give the same result. Hence R_1 and R_2 are convergent, we can generalize this on any connected replica R_i and R_j of the system, $i, j \in \mathbb{N}$.

3) *Lemma 4*: “If we've causal consistency in an asynchronous and unreliable distributed system, we can improve drastically the latency.”

(1) and (3) give:

$$L_G = \sigma \times (\lambda/c) \times MSG_OP \times IMPL_DTC + L_1 \quad (4)$$

RF can be ignored here; the messages are eventually delivered even if there's a failure or if messages are delayed for a finite duration. The causal consistency should be guaranteed at each node by his cache memory.

The number of messages across datacenters falls out in a spectacular way when we switch to the causal consistency. Consequently σ decreases drastically whether the number of message goes down; hence σ falls in whether we do not require higher than the causal consistency.

4) *Theorem 4.4*: “Any multi-datacentric distributed system which guarantees at most the causal consistency can provide the following properties: strong convergence, high availability and low latency.”

This theorem is a corollary of the corollary 1 and the lemmas 4 and 5.

E. Corollary 5.1 (SCOLCH theorem)

Based on the foregoing; we state the following theorem: “On multi-datacentric systems we have an exclusive choice between the strong consistency in a side; and low latency, strong convergence, and high availability in another side.”

IV. RELATED WORKS

A little over a decade ago the CAP theorem [10] shook the world of distributed systems and has significantly influenced the actors of the cloud computing. Accordingly the consistency on distributed databases at a large scale aroused considerable interest recent years. Obviously a series of papers published about the CAP conjecture by Brewer, Gilbert and Lynch, Abadi, Ramakrishnan, Shim, and Birman [1,9,11,17,24,26] have reignited the debate on the tradeoffs in the distributed systems. Additionally, in earlier 2000's, Brzezinski has published a good work on the causal consistency. Afterwards, Mahajan [23] and Shapiro [25] works have highlighted the concept of convergence. Mahajan et al. [23] proved also that the real-time causal consistency was insurmountable in unreliable distributed systems. Bailis proposed highly available transactions (HAT) as an alternative for ACID databases in the cloud within a series of publications [4, 5, 6]. He proposed also the PBS (Probabilistically Bounded Staleness) which could minimize the staleness of data and guarantee a limited latency [7]; he discussed also the limitations of eventual consistency [4]. CALM conjecture (Consistency And Logical Monotonicity) was proven by Alvaro who proposed also 'coordination-free' distributed modals [3]. Furthermore, Lloyd et al.[21] presented the design and implementation of COPS, a key-value store that delivers a causal consistency model with convergent conflict handling; he proposed also a scalable, geo-replicated storage system that guarantees low latency [22]. Finally, T. Kraska et al. [18] have proposed MDCC: an optimistic commit protocol for geo-replicated transactions.

V. CONCLUSION AND FUTURE WORK

The tradeoffs around the consistency property remain a nagging issue in multi-datacentric systems. In this paper we've lightened some basic concepts which are rarely explained by authors. Even if there are some theorems, the theoretical proofs are sorely lacking at this level, so we've given new theorems on these tradeoffs. Unlike the CAP theorem [10] which does not clearly take into account the latency and the convergence, our theorems prove that we can guarantee low latency, strong convergence and high availability in addition to the causal consistency which is proven to be better than the eventual consistency. In addition, our theorems prove that we cannot ensure the strong consistency without thereby sacrificing high availability and dealing with an unbridled growth of the latency.

In future work, it should be important clarify the fact that the very famous CAP theorem is not proved correctly and that CAP statement is out of context in actual distributed systems on the cloud.

ACKNOWLEDGMENT

We would like to thank the members of the database group of the doctoral school of mathematics and computer science of UCAD university of Dakar for their contributions to our research. We thank also the ICIT15 reviewers for their helpful feedback on this work.

REFERENCES

- [1] D. J. Abadi, Consistency Tradeoffs in Modern Distributed Database System Design, IEEE Computer Society, vol. 45, no. 2, pp. 37-42. 2012.
- [2] B., Alpern; F. B., Schneider. "Recognizing safety and liveness". Distributed Computing 2: pp. 117-126. (1987)
- [3] P. Alvaro, N. Conway, J. M. Hellerstein, and W. R. Marczak. Consistency analysis in Bloom: a CALM and collected approach. In CIDR 2011.
- [4] P. Bailis; A. Ghodsi. "Eventual Consistency Today: Limitations, Extensions, and Beyond". ACMQueue 11 : 20. April 2013.
- [5] P. Bailis, A. Ghodsi, J. M. Hellerstein, and I. Stoica. Bolt-on causal consistency. SIGMOD 2013.
- [6] P. Bailis, A. Davidson, A. Fekete, A. Ghodsi, J. M. Hellerstein, and I. Stoica. "Highly Available Transactions: Virtues and Limitations". In VLDB 2014.
- [7] Peter Bailis, Shivaram Venkataraman, Michael J. Franklin, Joseph M. Hellerstein, and Ion Stoica. "Quantifying Eventual Consistency with PBS". Communication of the ACM, vol. 57, n°8. August 2014.
- [8] P. Bernstein and S. Das. "Rethinking eventual consistency". In SIGMOD, 2013.
- [9] K. Birman and al., "Overcoming CAP with Consistent Soft-State Replication", IEEE Computer Society, vol. 45, Issue: 2 pp. 50- 58, February 2012.
- [10] E. Brewer, Towards Robust Distributed Systems, Portland, Oregon, Keynote at the ACM Symposium on Principles of Distributed Computing (PODC). July 2000.
- [11] E. Brewer, Pushing the CAP: Strategies for Consistency and Availability, IEEE Computer Society, pp. 23-29. February 2012.
- [12] J. Brzezinski, C. Sobaniec, and D. Wawrzyniak, "From session causality to causal consistency, in Proc. of 12th Euromicro Conf. on Parallel", Distributed and Network-Based Processing. Citeseer, pp. 152-158. 2004.
- [13] F. Chang et al, "BigTable: A Distributed Storage System for Structured Data", Seventh Symposium on Operating System Design and Implementation, November 2006.
- [14] B. Cooper et al. PNUTS: Yahoo!'s hosted data serving platform. In VLDB 2008.
- [15] B. DeCandia and al., "Dynamo: Amazon's Highly Available Key-Value Store", Proceedings 21st ACM SIGOPS Symposium on B. F. Cooper, PNUTS: Yahoo!'s Hosted Data Serving Platform, Proc. VLDB Endowment (VLDB 08), pp. 1277-1288. 2008.
- [16] P. Deutsch. "The eight fallacies of distributed computing." <http://tinyurl.com/c6vvtzg>, 1994.
- [17] S. Gilbert and N. Lynch, Perspectives on the CAP theorem, IEEE Computer Society, vol. 45, no. 2, pp. 30-36. 2012.
- [18] T. Kraska, G. Pang, M. Franklin, and S. Madden. "Mdcc: Multi-data center consistency". In Eurosys 2013.
- [19] A. Lakshman and P. Malik, Cassandra: a decentralized structured storage system, ACM SIGOPS Operating Systems Review, vol. 44, , pp. 35-40. 2010.
- [20] L. Lamport, L.. "Proving the Correctness of Multiprocess Programs". IEEE Transactions on Software Engineering : (1977) march, pp. 125-143.
- [21] W. Lloyd, M. J. Freedman, M. Kaminsky, and D. G. Andersen. "Don't Settle for Eventual: Scalable Causal Consistency for Wide-Area Storage with COPS". In SOSP 2011.
- [22] W. Lloyd, M. J. Freedman, M. Kaminsky, and D. G. Andersen. Stronger semantics for low-latency georeplicated storage. In NSDI 2013.
- [23] P. Mahajan , L. Alvisi , M. Dahlin. Consistency, Availability, and Convergence. Department of Computer Science, University of Texas at Austin. Technical Report (UTCS TR-11-22).2012.
- [24] R. Ramakrishnan, CAP and Cloud Data Management, IEEE Computer Society, vol. 45, Issue: 2 pp. 43 -49, February 2012.
- [25] M. Shapiro, Convergent and Commutative Replicated Data Types, Bulletin of the EATCS, no. 104, pp. 67-88. June 2011.
- [26] S. S. Y. Shim, CAP theorem's growing impact, IEEE Computer Society, vol. 45 , Issue: 2 pp 20 -21, February 2012.
- [27] M. Stonebraker, Errors in Database Systems, Eventual Consistency, and the CAP Theorem, blog, Comm. ACM, <http://cacm.acm.org/blogs/blog-cacm/83396-errors-in-database-systems-eventualconsistency-and-the-cap-theorem>. 2010.
- [28] M. Stonebraker, Clarifications on the CAP Theorem and Data-Related Errors, VoltDB blog, <http://blog.voltdb.com/clarifications-cap-theorem-and-data-related-error>. 2010.
- [29] W. Vogels, "Eventually Consistent", ACM Queue, vol. 6, n° 6, pp. 14-19. 2008.

Computer Networks and Communications

Analyzing Optimal Setting Of Reference Point Group Mobility Model Using DSR Protocol In MANETS

Nasser Ali Husieen

Department of Computer Science, College of Education
Wasit University
Kut, Iraq
Email: dr.naseer.alquraishi@gmail.com

Mohammad M. Rasheed

Information Technology Directorate
Ministry of Science and Technology
Baghdad, Iraq
Email: mohmadmhr@yahoo.com

Abstract— Mobile Ad-hoc Network (MANET) is a self-configuring wireless network. The nodes can configure themselves to be in any arbitrary topology. It is essential that mobility models used in simulating different scenarios must emulate closely the real scenario in order to assess the MANET performance as accurately as possible. In this paper, the authors have studied the effect of different maximum pause times and maximum node speeds on different performance metrics in order to arrive at optimal settings for these two attributes under the Reference Point Group Mobility model for the DSR protocol. In addition, this work is part of an ongoing research on link failures in DSR protocol. Thus, the performance of the DSR protocol under the Reference Point Group Mobility (RPGM) model in terms of different pause times, node speeds, number of nodes and number of source connections were evaluated. The simulation results show that the maximum pause time and the maximum speed have direct impacts on the performance parameters such as packet delivery ratio, routing overhead, average end-to-end delay, normalized routing load and packet drop under the Reference Point Group Mobility model.

Keywords— MANET, Mobility Models, Routing Protocols, DSR Protocol, RPGM

I. INTRODUCTION

Mobile ad hoc network is a collection of wireless nodes that allows devices to communicate with each other without help of an existing infrastructure [1]. A MANET is a self configuring and a self-organizing such that it can create an arbitrary topology temporarily for short while in which mobile nodes work as both routers and end nodes. MANET based applications range from military applications connecting soldiers in a battlefield, social networks for communication during emergencies such as natural disasters to personal area networks. New applications such as telemedicine, weather reporting and dissemination and disaster information dissemination are emerging in the recent times exploiting the new developments and the advantages of MANETs compared to the traditional infrastructure based networks. The above mentioned developments will increase the size and reach of MANETs to thousands of nodes which is hitherto limited to few nodes in both military and civil application domains.

One of the main challenges in setting up and managing a dynamic mobile ad hoc network is routing. Routing protocols detect the optimum path between the source node and destination node in a complex network of nodes and deliver

the data packets between those nodes in an efficient manner. In mobile ad hoc networks, high mobility, limited computing capability and low bandwidth associated with nodes make routing of data one of the most challenging tasks. Researchers have proposed different routing protocols to address these issues. These protocols can be categorized into two main groups. They are namely on demand or reactive protocols and periodic or proactive protocols. Protocols such as Dynamic Source Routing (DSR) [2], Ad-hoc On-Demand Distance Vector Routing (AODV) [3] and Temporally ordered routing algorithm (TORA) [4] coming under the reactive protocols establish the route only prior to sending the packets and maintain the route only when it is needed. The proactive protocols such as Destination-Sequenced Distance Vector (DSDV) [5] on other hand periodically exchange routing information to maintain the routes continuously. In general, reactive protocols have been shown to outperform the proactive protocols due to their reduced overhead and ability to react quickly when routes change [6-9]. This motivated us to investigate more into on demand protocols. One of the most important and difficult tasks in simulating mobile ad hoc networks is the definition of a mobility model to represent real world scenarios. In the real world, mobile networks encounter various situations including intense mobility creating heavy

uncertainties especially in disaster situations. Hence taking measurements under real world conditions is an almost impossibility. The alternative to real world measurements is to simulate the different situations using a computer software represent the real world situation as closely as possible and take the measurements. Presently in the mobile ad hoc environments, the following mobility models have been used; Random Way Point, Manhattan Grid, Gauss Markov mobility model and Reference Point Group Mobility (RPGM) model [10]. In this paper, we use the RPGM model implemented in the Network Simulator 2 (NS2) version 2.34 under Linux to study the effect of pause time and node speed on the DSR protocol. Different scenarios were simulated and the performance of the network was critically evaluated with special reference to these parameters. The main objective of the analysis was to determine the optimal settings for these two parameters.

The rest of the paper is organized as follows: Section 1 provides an introduction to the study. Sections 2 and 3 discuss RPGM model in details and related work respectively. Details of the simulation environment setup and results are presented in Sections 4 and 5. Finally Sections 6 concludes the paper along with suggestions for future work.

II. DYNAMIC SOURCE ROUTING PROTOCOL

Dynamic Source Routing (DSR) protocol is a simple reactive routing protocol developed by Johnson at the Carnegie Mellon University in 1996. The key feature of the DSR protocol is based on the source routing where full or partial route is specified by the sender [11]. In mobile ad hoc networks, the source routing technique provides several advantages, including flexibility, simplicity and correctness [12-14]. When a node wants to send data packets to another node, the intermediate nodes forwarding the packets towards the destination need not maintain up to date routing information as the data packet will contain information on how to forward the packet. DSR protocol is made up of two mechanisms namely, Route Discovery and Route Maintenance. DSR uses route cache to store the routes to other nodes. The main advantage of using route cache is it speeds up route discovery and reduces the propagation of route request packets. DSR can also provide interconnection between wireless devices with multiple network interfaces. This is vital in strategic communications as nodes in the military require communicating with different devices at different ranges

Fig. 2 shows the route maintenance mechanism of DSR protocol. In this mechanism, DSR used two types of packets: ACKs packet is used for correction and verification operation of the routes, and second type of packets is called Route Error Packet (RRER). RRER is generated when there is link failure occurs between intermediate nodes and next hop due to battery usage, hard medium contention, and the node mobility that is leading to loss the packets as shows in the figure below when the link fails between the intermediate node I and destination node D.

Fig. 1 shows the route discovery mechanisms. In this figure, when a source node (A) wants to send data packets to the destination node (J). First, it checks if there is a route to the destination in the route cache, route request will be discarded. Otherwise, source node (A) broadcasts Route Request Packet (RREQ) to the neighbour nodes (B, C, D) within wireless transmission range of a node (A). The route request identifies the destination node J to which a route is needed. If the neighbour nodes (B, C, D) are not the target, locally these neighbour nodes rebroadcast RREQ to the next hop after adding its own address to a list of nodes in the route cache. Each RREQ packet contains source address, destination address, request ID, and route record. When a route request packet arrives at the target node (J), the destination node (J) returns a Route Reply Packet (RREP) along with the reverse of a recorded path to the source node (A), which is (A, D, G, J). When a source node (A) receives more than one RREP for a given destination, it selects the first route that receives RREQ in order to reduce the time for route discovery packet. DSR protocol is made up of two mechanisms namely, Route Discovery and Route Maintenance. DSR uses route cache to store the routes to other nodes. The main advantage of using route cache is it speeds up route discovery and reduces the propagation of route request packets.

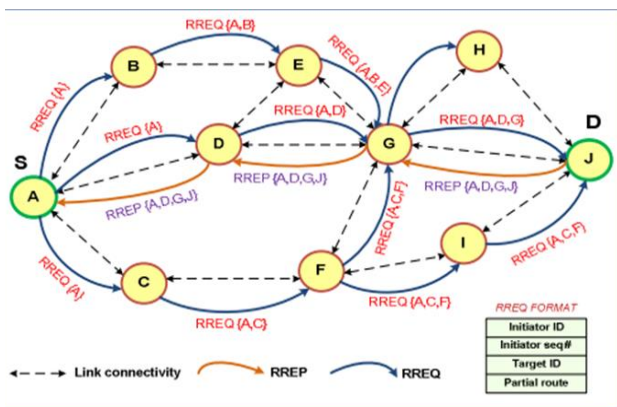


Fig.1. Route Discovery Mechanism

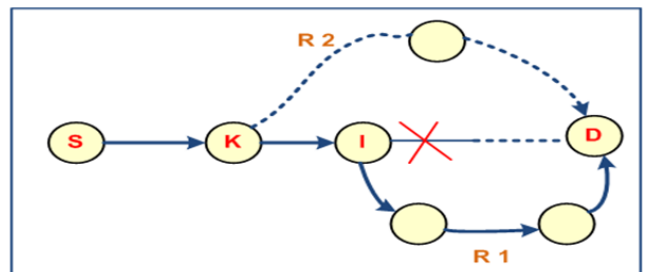


Fig.2. Route Maintenance Mechanism

III. REFERENCE POINT GROUP MOBILITY MODEL

The Reference Point Group Mobility (RPGM) model was proposed by Hong [15]. In this model, all nodes work as a group and the nodes of the group moves as a single entity to achieve different tasks. Each group has the logical centre called the group leader. The path of the group in its entirety is represented by the locus of the centre. Each node in the group has its own reference point for communicating with other nodes. The reference point of the node follows the group movement; the real location of the node can be determined by its reference point plus a random motion vector that denotes its stability from the reference point. Reference point Group mobility is adapted for several applications such as a battlefield situation where a number of soldiers move together in a group, disaster recovery and convention scenarios. According to Hong's report, the RPGM outperforms Random Way Point model in case of link failures due to the inherent characteristic of spatial dependency between nodes. The RPGM model incurs fewer link breakages and achieves better performance for various routing protocols compared to Random Way Point model. The functions of the group leader and group members are as follows:

A. Group Leader

V_{group}^t It provides the general motion movement of the whole group. Each member of this group moves away from this group motion. The motion vector V_{group}^t can be arbitrarily selected or carefully designed based on several predefined paths.

B. Group members

The group members' movement is heavily affected by its group leader's movement. Each mobile node is assigned with a reference point that follows the group movement. With respect to this predefined reference point, every mobile node might be arbitrarily located in the neighborhood. Formally, the motion vector of the group members i , at the time t , V_i^t can be defined as:

$$V_i^t = V_{group}^t + RM_i^t \quad (1)$$

Where RM_i^t is the random motion vector representing the deviation of the group member i from its reference point. The vector RM_i^t is free identically distributed random procedure whose duration is uniformly distributed in the interval $[0, r_{max}]$, where r_{max} is maximum acceptable distance and whose path is uniformly distributed in the interval $[0, 2\pi]$.

Fig. 3 illustrates the Reference Point Group Mobility model with the group leader represented in green and the members represented in red and yellow respectively. V_{group}^t is the motion vector of the group leader and the whole group.

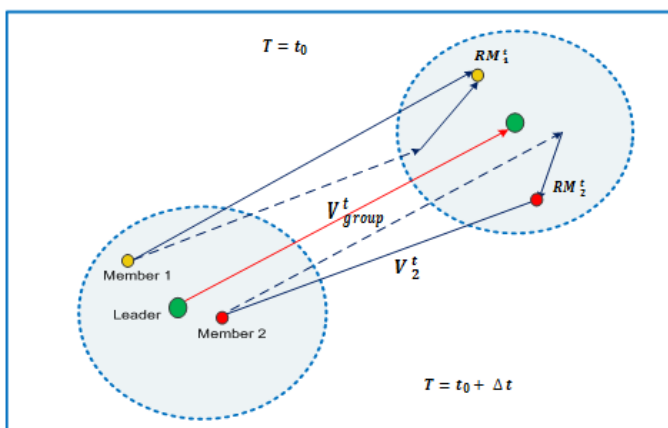


Fig.3. Node Movement in RPGM Model

With appropriate selection of the predefined paths for the group leader and other parameters, the RPGM model can emulate a variety of mobility behaviours. The RPGM model is thus able to represent various mobility scenarios, such as:

- *In-Place Mobility Model:* In this model, the whole field is divided into adjacent regions. Each region is totally occupied by a single group. An example of this model is battlefield communication.
- *Overlap Mobility Model:* In this model, different groups with different tasks move on the same field in an overlapping style. An example of this model is Disaster relief.
- *Convention Mobility Model:* In this model, the area is divided into few regions and several groups are permitted to move between the regions. An example of this model is a conference.

In RPGM model, the vector RM_i indirectly determines how far the group members deviate from their leader. The movement can be characterized as follows:

$$|V_{member}(t)| = |V_{leader}(t)| + random() * SDR * max_speed$$

$$\theta_{member}(t) = \theta_{leader}(t) + random() * ADR * max_angle \quad (2)$$

Where $0 < SDR, ADR < 1$,

SDR is the Speed Deviation Ratio and

ADR is the Angle Deviation Ratio

SDR and ADR are used to control the deviation of the velocity in terms of both magnitude and direction of group members from that of the leader. Different mobility scenarios can be generated by adjusting these two parameters.

IV. RELATED WORK

Different mobility models have different impacts on the performance of mobile ad hoc routing protocols [16]. Different protocols have different metrics that capture interesting mobility characteristics like spatial and temporal dependence and geographic restrictions. Hong et al [15] introduced a mobility model called the Mobility Vector (MV) model and compared the performance of different routing protocols such as DSR, AODV and FSR against other mobility models including Random Way Point, RPGM and Random Walk. Packet delivery ratio and link Up/Down were measured using simulation varying the average speed and transmission range respectively. Hence it is important to select the right mobility model to represent the real scenario under

consideration. Geetha and Gopinath compared the performance of two on demand routing protocols, namely AODV and DSR under different mobility models such as Random Way Point and Reference Point Group Mobility in order to characterize the two routing protocols under the different mobility models [17]. Simulation parameters computed included fixed pause time at 25 s, five low traffic source and speeds up to 20 m/s. In this study, the authors concentrate on different pause times and maximum speed with increased number of source connection along with increased node density in order to find the optimal setting of pause time and node speed for the DSR protocol in the Reference Point Group Mobility.

Kioumourtzis evaluated the performance of OLSR, AODV and DSR three reactive protocols under Manhattan-Grid, and Reference Point Group mobility models [18]. The main objective of his work was to compare the performance of each protocol under the two models in order to understand the limitations of the protocols. The simulations were carried out under different conditions. The number of nodes was varied up to 90, the number of connections was increased up to 40, the packet rate was also increased and the nodes' speeds of movement were fixed at 5, 10, 15 and 20 m/s along with a pause time of 5 s. However, this paper presents the performance of the DSR protocol with increased node density, increased CBR traffic and various nodes' mobility velocities up to 80 m/s with different pause times up to 40 s in order to select the optimal setting for the DSR protocol under Reference Point Group Mobility model.

Agrawal, Tiwari and Vyas have evaluated the AODV protocol under four different mobility models, namely the Manhattan-Grid, Markov-Grid, Random Way Point and Reference Point Group Mobility model [19]. Their objective

was to select a suitable model for AODV protocol. The metrics used by them were packet delivery ratio and delay. Their simulation result shows that the AODV performs well with RPGM model in terms of packet delivery ratio and end to end delay. The simulation parameters used were fixed pause time of 10 s, different velocities of 5, 10, 15, 20, 25 m/s and increased network load of 4, 8, 12 and 16 packets per second. The work presented by this paper focuses on velocities up to 80 m/s with different pause times up to 40 with the objective of selecting the average optimal setting for these two parameters.

V. SIMULATION ENVIRONMENT

The Network Simulator 2 (NS2) version 2.34 installed on Centos Linux operation system was used as the simulation tool in this work. In NS2 the node movement has to be defined in a OTCL script or be imported from an external file. In this project, the mobility scenarios were created using the Bonn Motion version 1.4 a Java software tool specifically designed for this purpose [20]. Bonn Motion was developed by the Communication Systems Group at the Institute of Computer Science 4 of the University of Bonn, Germany. This tool can generate the most common mobility models such as Random Way Point, Gauss-Markov and Manhattan-Grid models. In this project, the Reference Point Group mobility (RPGM) model has been used as a movement model and the `cbrgen.tcl` generator tool which is located under the directory `indep-utills/cmu-scen-gen` has been used to generate random source traffic. The traffic pattern used is CBR. Fig. 4 shows the simulation methodology adopted in this work. Each scenario was simulated for 200 seconds within a simulated rectangular geographical area of 1000 m x 500 m. Tables 1 and 2 lists the rest of simulation parameters.

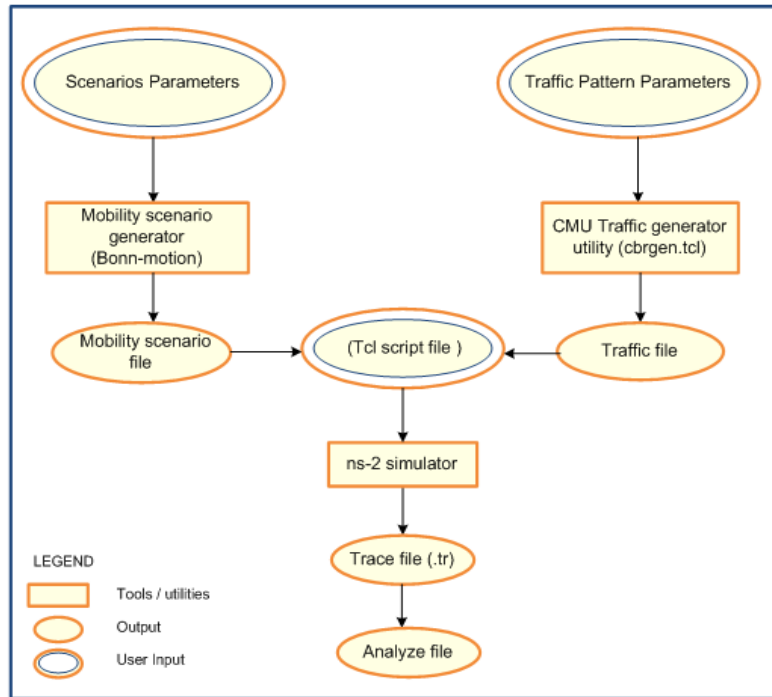


Fig 4. Simulation Methodology

TABLE 1: Simulation Parameters

Parameters	Value
Simulation Time	200 s
Number of nodes	10, 20, 40, 80
Number of connections	4, 8, 30, 40
Maximum Pause Time	5, 10, 20, 30, 40
Simulation Area	1000 x 500 m
Minimum node speed	0 m/s
Maximum node speed	20,40,60,80 m/s
Mobility Model	RPGM
Routing Protocol	DSR
Traffic Type	Constant Bit Rate (CBR)
Packet size	512 bytes
Distribution of nodes	5 groups
Probability of group change	0.05
Maximum distance to group centre	100 m

Standard deviation	2.0
MAC Type	802.11

According to the data listed in Table 1, there are eight main scenarios comprising four main scenarios of different speeds and another four main scenarios of different pause times. In the case of first main scenarios, for each speed was tested under four sub-scenarios by changing different metrics at a time making the total number of scenarios under this category to be 16. Under each pause time, five sub-scenarios were created using different additional parameters making the total number of scenarios under this category to be 20. Thus, the total number of scenarios tested in the experiment is 36. Since the performance of ad hoc routing protocols is sensitive to the movement pattern, scenario files were generated with 50 different movement patterns representing 10 movement

```

    ./bm -f scenatio1 -b RPGM -n 80 -d 200 -x 1000 -y
    500 -h 20.0 -l 0.0 -p 20.0 -a 5.0 -c 0.05 -r 100 -s
    2.0
    
```

patterns per pause time.

Table 2 lists the parameters used in generating the RPGM mobility model in Bonn Motion.

The following command will generate the Reference Point Group Mobility scenario.

	centre
-s	Group size standard deviation

TABLE 2: RPGM Parameters in Bonn Motion

Parameters	Explanation
-n	Number of mobile nodes
-d	Simulation duration time
-x	Simulation area width
-y	Simulation area height
-c	Group change Probability
-l	Lowest velocity
-h	Highest velocity
-p	Pause time
-a	Average number of nodes per group
-r	Maximum distance to group

After the generation of scenario1, the following command should be typed to transform the scenario1 into a file that can be read by ns-2. Fig. 5 shows the creation of clusters as shown in the NAM console at completion of the simulation.

```
./bm NSFile -f scenario1
```

VI. SIMULATION RESULTS

Four different scenarios were considered with different pause times. Each scenario contained five sub scenarios along with four different scenarios for different node speeds containing four sub scenarios each. The simulation for five times for each scenario and the average of the results were computed. The packets delivery ratio, average end to end delay, routing overhead, normalized routing load and packet drop were used as metrics in evaluating performance.

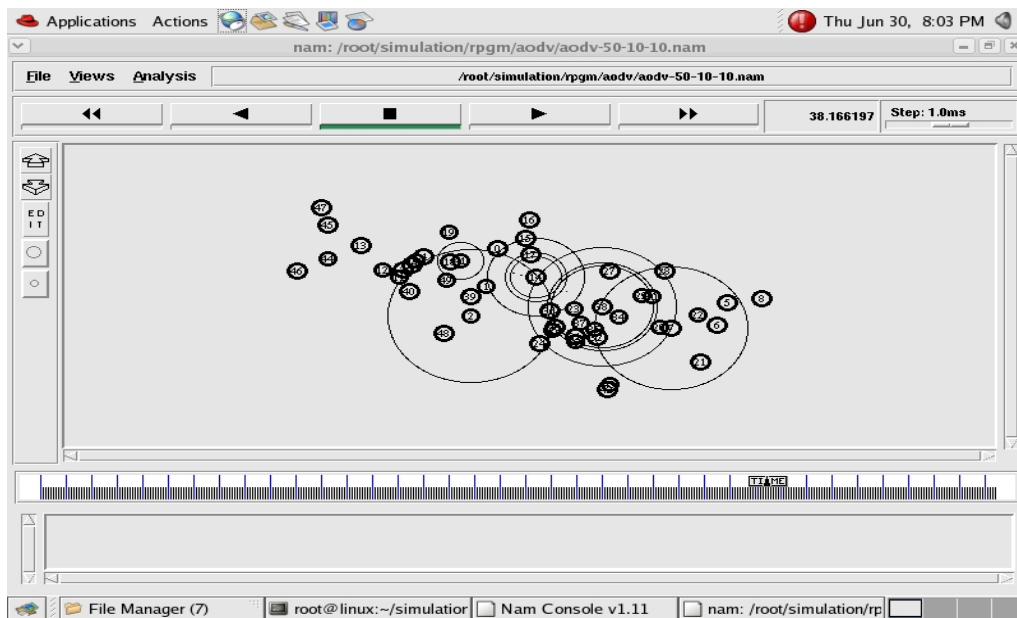


Fig. 5. Creation of Clusters with the RPGM model

A. Packet Delivery Ratio

Packet Delivery Ratio (PDR) is defined as the fraction of the number packets received at the destination to the number of packets originated by the source application. PDR describes the loss rate seen by the transport layer protocols. This will affect the overall network throughput.

The PDR as a percentage is given by Formula (3).

$$\left[PDR\% = \frac{\sum_1^n DATA_{rcv}}{\sum_1^n DATA_{sent}} \times 100 \right] \quad (3)$$

Fig. 6. (A) shows the PDR for different pause times. The investigation was carried at pause times 5, 10, 20, 30 and 40 s and the PDR at these pause times were measured in order to select the best pause time. From Fig. 6 (A), it can be seen that the DSR protocol performs well when the pause time is 20 s under all the four scenarios. The results show that the highest PDR with the maximum number of packets sent and received at the pause time 20 s, is 100.00%, 100.00%, 99.35% and 99.96%. After selecting the pause time to be fixed at 20 s, different maximum speed of 20, 40, 60 and 80 m/s were investigated in order to select the optimal node speed in the RPGM model.

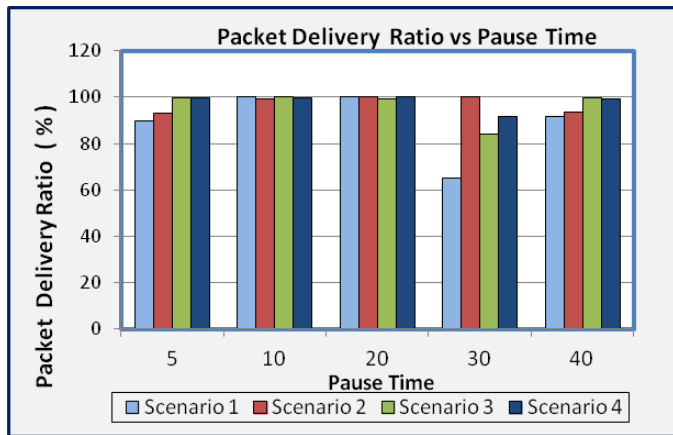


Fig. 6 (A). PDF vs. Pause time

Figure 6(B) shows the effect of node speed on the DSR protocol. The results show that the highest PDR is achieved when maximum speed is 20 m/s or 72 km/h. The PDR at 20 m/s were 100.00%, 99.96%, 99.97% and 99.88% respectively. Whenever the node speed was increased, the PDR dropped due to the reason that the group leader and the members were moving very fast during the packet delivery process. This causes packets to drop leading to reduced PDR. It was also observed that when a mobile node moves fast and the pause time is small; the topology is likely to be in a highly dynamic condition. Hence, the optimal setting for node speed and the pause time are 20 m/s and 20 s respectively.

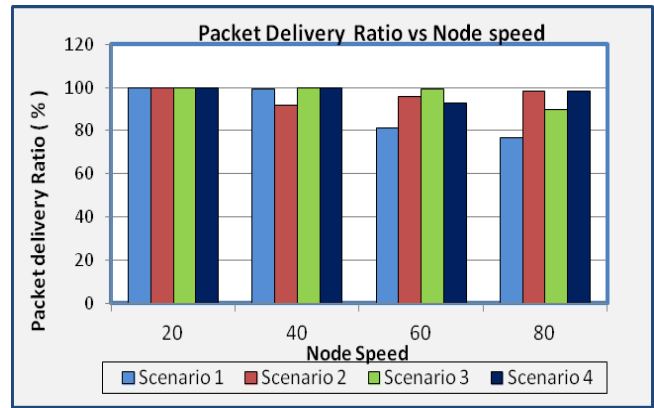


Fig. 6 (B). PDF vs. Node speed

B. Average End-to-End Delay

Average end-to-end delay is defined as the average time taken by a packet to reach the destination from the source. The average end-to-end delay includes all the types of delays such as transmission delay, propagation delay, processing delay and interface queuing delay. The average end-to-end delay between two nodes is computed using Formula (4).

$$\left[Avg_Delay = \frac{\sum_1^n DATA_{sent} - DATA_{rcv}}{\sum_1^n DATA_{rcv}} \right] \quad (4)$$

The delay is affected by higher CBR packet generation rate at the source as well. When packets are generated at high rates, the buffer at the source becomes full resulting in longer queuing delays at the source node. Figures 8(A) and (B) show the effect of maximum pause time and speed on average end-to-end delay respectively. Fig. 7(A) shows that the average end to end delay is reduced when the pause time is 20 s in the 2nd and 4th scenarios. The average end-to-end delays were 30.56 s and 15.82 s respectively. The average end to end delays in the 1st and 3rd cases at 10 s pause time were 6.62 s and 11.66 s respectively. The reason for this discrepancy is the increased time consumption for route discovery at 20 s pause time due to buffer overflow in the 1st and 3rd scenarios.

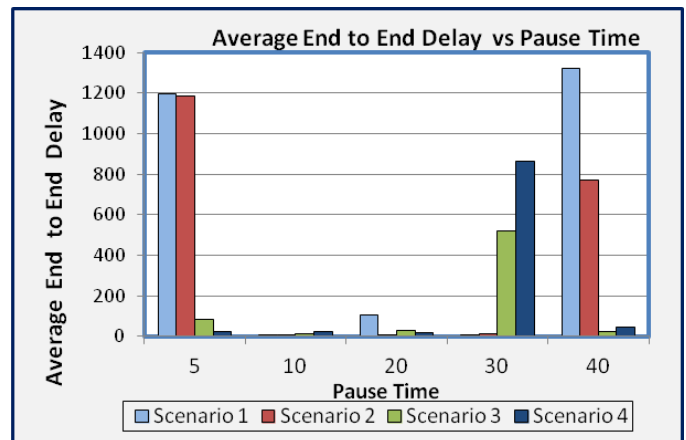


Fig 7(A). Delay vs. Pause Time

Fig. 7(B) shows the effect of different maximum speeds on the average of end-to-end delay. From Fig. 7(B) it can be seen that

the average end-to-end delay is decreased when maximum speed is 20 m/s. An increase in maximum speed also results in an increase in the average end-to-end delay. This is due to the reason that when a node moves fast, it causes link failures leading to new route discovery processes. New route discovery processes would create additional delays increasing the overall end-to-end delay. Hence it can be concluded that the optimum settings for the maximum pause time and the speed are 20 s and 20 m/s respectively.

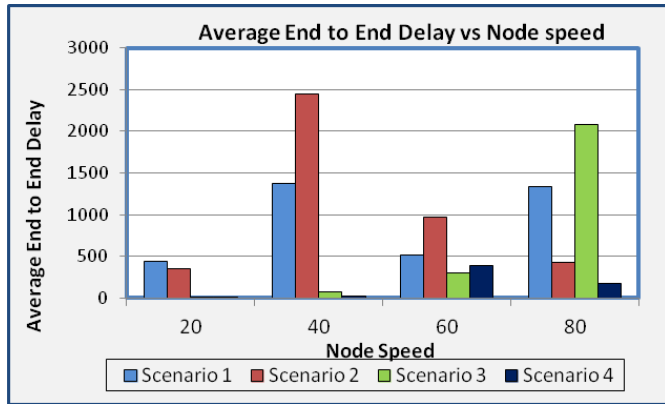


Fig. 7 (B). Delay vs. Node Speed

C. Routing Overhead

Overhead is an important issue as higher overheads reduce the overall network resources utilization. Hence it is important to reduce the overhead in a network as much as possible. Fig. 8 (A) and (B) show the routing overhead for different pause times. From Fig. 9(A), it can be seen that different pause times affect the routing overhead under all the four scenarios. Fig. 8(A) shows that the overhead is the minimum when maximum pause time is 20 s, except for the 3rd scenario with total number of nodes to be equal to 40 with 30 source connections and the maximum pause time to be 10 s. The 3rd scenario has the optimum setup at the maximum pause time equal to 10 s due to the lower number of packets transmitted. In addition, DSR protocol uses a route cache to reduce the number of route discovery processes during the establishment and transmission of packets.

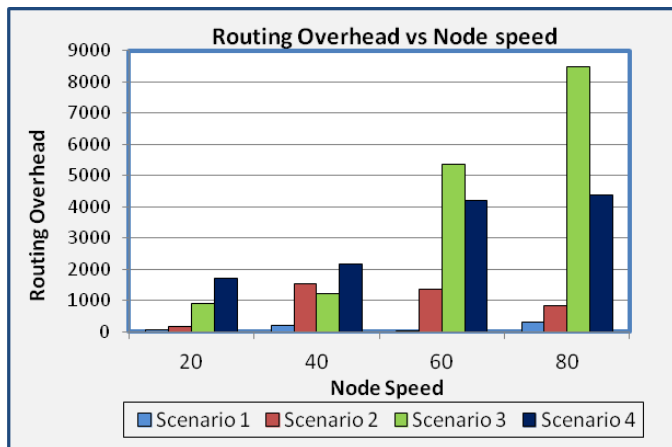


Fig. 8 (A). Overhead vs. Pause time

Fig. 8(B) shows that different maximum speeds affect the routing overhead. It can be seen that the routing overhead is the minimum at the maximum speed is at 20 m/s in all scenarios. Whenever the maximum speed of a node is increased, the overhead is also increased. Hence, it can be concluded that the average optimal settings for maximum pause time and speed are 20 s and 20 m/s for minimizing the routing overhead.

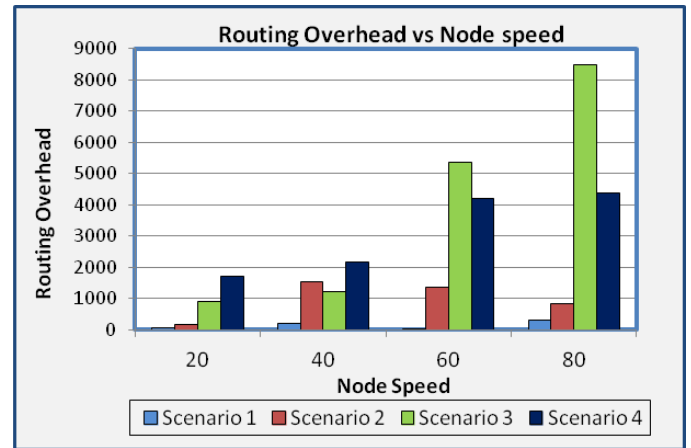


Fig. 8 (B). Overhead vs. Node speed

D. Normalized Routing Load (NRL)

The Normalized Routing Load (NRL) is an estimate of how efficient a routing protocol is. The number of routing packets sent per data packet is an indication of how well the protocol maintains the routing information updated. The higher the normalized routing load, the higher the overhead of routing packets is and consequently the lower the efficiency of the protocol. Fig. 9(A) and (B) show the results for the NRL against different maximum pause times and maximum node speed respectively. Fig. 9(A) shows that the DSR protocol performs best when maximum pause time is 20 s except for the 3rd scenario, where the pause time is 10 s. This is due to the reason that the NRL is directly proportional to the overhead and the packets sent. As shown in Fig. 10(A), the NRL is reduced when the maximum pause time is 10 s.

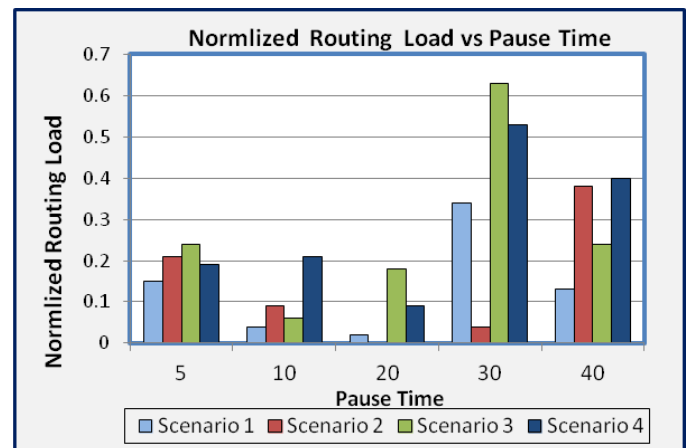


Fig. 9 (A). NRL vs. Pause time

Fig. 9(B) shows the effect of the different maximum node speeds on the NRL. Fig. 9(B) shows that the DSR protocol performs best when maximum node speed is 20 m/s resulting in the lowest NRL value. Whenever the maximum node speed is increased, the NRL is also increased leading to lower efficiency. Hence, it can be concluded that the optimal setting for the maximum pause time and the maximum node speed are 20 s and 20 m/s respectively for efficient NRLs.

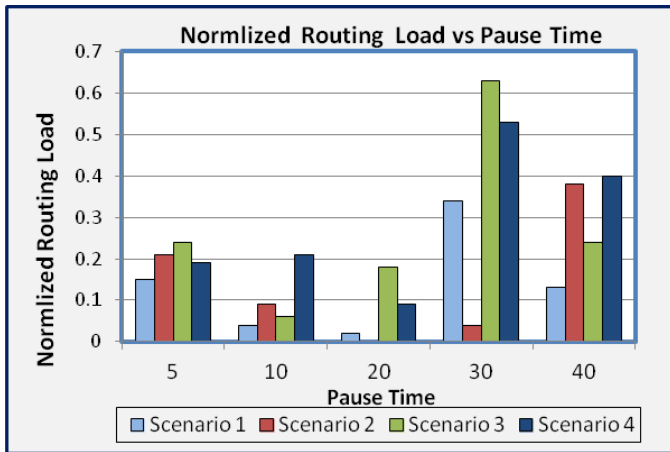


Fig. 9 (B). NRL vs. Node speed

VII. CONCLUSIONS

In this paper, the authors have presented the work of evaluating the performance of the DSR protocol under Reference Point Group Mobility model with respect to the effects of the maximum pause time and maximum speed under different scenarios. The simulation parameters included 36 different scenarios of which 20 scenarios are for different maximum pause times and 16 scenarios are for different maximum node speeds. Simulation results show that the maximum pause time and speed have a direct impact on the performance of the DSR protocol. The Reference Point Group Mobility model has two constraints. One is that the movement of nodes should follow that of the group leader and the other one is that each group leader has the group movement limited to certain speed limits. The simulation results show that when the mobility is high, the possibility of route caches becoming stale is also high resulting in frequent link failures. When a route discovery is initiated, the large number of replies received in response is associated with high MAC overhead causing increased interference with data traffic. Hence, the cache staleness and high MAC overhead together result in significant degradation of performance in DSR in high mobility scenarios. Furthermore, simulation results show that there are four main factors that can affect the DSR protocol. They are namely, the maximum node speed, the maximum pause time, the number of connections, and the number of nodes. From the simulation results, it was observed that the optimal settings for the maximum pause time and the maximum node speed are 20 s and 20 m/s respectively.

REFERENCES

- [1] J. Jain, M. Fatima, and R. Gupta, "Overview Challenges of Routing Protocol and Mac Layer in Mobile Ad-hoc Network," *Journal of Theoretical and Applied Information Technology* vol. 8, pp. 6-12, 2009.
- [2] C. E. Perkins and E. M. Royer, "Ad-Hoc on-Demand Distance Vector Routing," in *The Second IEEE Workshop on Mobile Computing Systems and Applications*, 1999, pp. 90-100.
- [3] M. S. Corson and A. Ephremides, "A Distributed Routing Algorithm for Mobile Wireless Networks," *Journal of Wireless Networks*, vol. 1, pp. 61-81, 1995.
- [4] C. E. Perkins and P. Bhagwat, "Highly Dynamic Destination-Sequenced Distance-Vector Routing (DSDV) for Mobile Computers," in *SIGCOMM '94 Conference on Communications Architectures, Protocols and Applications*, 1994, pp. 234-244.
- [5] J. Broch, D. A. Maltz, D. B. Johnson, Y. Hu, and J. G. Jetcheva, "A Performance Comparison of Multi-Hop Wireless Ad Hoc Network Routing Protocols," in *Fourth Annual ACM/IEEE International Conference on Mobile Computing and Networking* 1998, pp. 85-97.
- [6] P. Johansson, T. Larsson, N. Hedman, B. Mielczarek, and M. Degermark, "Scenario-based Performance Analysis of Routing Protocols for Mobile Ad-hoc Networks," *Fifth Annual ACM/IEEE International Conference on Mobile Computing and Networking*, pp. 195-206, 1999.
- [7] D. A. Maltz, J. Broch, J. Jetcheva, and D. B. Johnson, "The Effects of On-Demand Behaviour in Routing Protocols for Multi-Hop Wireless Ad Hoc Networks," *IEEE Journal on Selected Areas in Communications*, Special Issue on "Wireless Ad Hoc Networks", vol. 1439-1453, pp. 1439-1453, 1999.
- [8] S. R. Das, C. E. Perkins, and E. M. Royer, "Performance Comparison of Two On-demand Routing Protocols for Ad Hoc Networks," in *Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies 2000*, pp. 3-12.
- [9] P. Prabhakaran and R. Sankar, "Impact of Realistic Mobility Models on Wireless Networks Performance," in *Wireless and Mobile Computing, Networking and Communications*, 2006, pp. 329-334.
- [10] D. B. Johnson, "Routing in Ad Hoc Networks of Mobile Hosts," in *IEEE Workshop on Mobile Computing Systems and Applications*, 1994, pp. 158-163.
- [11] D. B. Johnson and D. A. Maltz, "Dynamic Source Routing in Ad Hoc Wireless Networks," in *Mobile Computing*, ed, 1996, pp. 153-181.
- [12] D. B. Johnson, D. A. Maltz, and Y. Hu. (2004). Dynamic Source Routing Protocol for Mobile Ad-hoc Networks (DSR). Available: <http://www.ietf.org/internet-drafts/draft-ietf-manet-dsr-10.txt>
- [13] D. B. Johnson, D. A. Maltz, Y.-C. Hu, and J. G. Jetcheva. (2001). The Dynamic Source Routing Protocol for Mobile Ad Hoc Networks, . Available: Internet- Draft, draft-ietf-manet-dsr-05.tx
- [14] D. B. Johnson, D. A. Maltz, and J. Broch, "The Dynamic Source Routing Protocol for Multi-hop Wireless Ad Hoc Networks," in *Ad Hoc Networking*, ed, 2001, pp. 139-172.
- [15] X. Hong, M. Gerla, G. Pei, and C. C. Chiang, "A Group Mobility Model for Ad Hoc Wireless Networks," in *2nd ACM International Workshop on Modelling, Analysis and Simulation of Wireless and Mobile Systems*, 1999, pp. 53-60.
- [16] F. Bai, N. Sadagopan, and A. Helmy, "Important: a Framework to Systematically Analyze the Impact of Mobility on Performance of Routing Protocols for Ad Hoc Networks," in *22nd Annual Joint Conference on the IEEE Computer and Communications Societies USA*, 2003, pp. 825-835.
- [17] G. Jayakumar and G. Gopinath, "Performance Comparison of Two On-demand Routing Protocols for Ad-hoc Networks based on Random

- Waypoint Mobility Model," Journal of Applied Sciences American, vol. 6, pp. 659-664, 2008.
- [18] G. Kioumourtzis, "Simulation and Evaluation of Routing Protocols for Mobile Ad Hoc Networks," MSc. Thesis, Naval Postgraduate School, California, 2005.
- [19] C. P. Agrawal, M. K. Tiwari, and O. P. Vyas, "Evaluation of AODV Protocol for Varying Mobility Models of MANET for Ubiquitous Computing," in Proceeding of the Third International Conference on Convergence and Hybrid Information Technology, 2005.
- [20] Bonn Motion Mobility Generator
www.informatik.uni-bonn.de/IV/BonnMotion.

Quantum Information Technology: Novel Way for Increase of Sensory Systems Capability

Paata J. Kervalishvili

Department of Engineering Physics, Georgian Technical University, Tbilisi, Georgia

Abstract – In the last decade quantum information theory and technology evolve and show their great potential. There are a set of problems for which it's more efficient and even not possible with classical communication to solve than with quantum equivalent. The best known example is Quantum Key Distribution (QKD), though there are quantum non-locality (entanglement), quantum teleportation, communication complexity and many more. The quantum information technologies permit by using quantum representation of data, to collect much bigger, more varied and precise information, as well as quantum data bank creation, which can be effectively treated by usage of relevant quantum algorithms. For creation of quantum databases, two methods will be dealt with: One is based on quantum numbers usage for processing various parametrical values (attributes of data bank); next, the database will be presented as its quantum model. On the basis of quantum information technology approach the new methods of possible improvement of nano micro sensory systems effectiveness is discussing in the recent paper. Multiparametral and multifunctional nature of sensors and their networks was taking into account. Nano micro sensor systems integrate and interface multiple core technologies and related devices to implement a variety of functions. They can be implemented through scalable homogeneous, or heterogeneous hardware integration technologies, in order to advance the miniaturisation, functionality and reliability of the sensor, processor, actuator and communication functions. Power autonomy (consumption and supply) is a common issue. In the medium term, there is growing industrial interest to integrate nanosensors in smart (intelligent) microsystems, mainly due to an increase in sensitivity, device simplification and associated cost reduction.

Keywords – quantum information, sensory systems, swarm intelligence, qubit, data collection and processing

I. INTRODUCTION

Monitoring natural uncertain environment parameters is a complex task of great importance in many areas. The origin of the difficulty lies in the environment's dynamism, arguably representative of real world problems, which consists of a number of peaks of changing width and height and in diffuse processes [1,2].

For technological monitoring of environmental safety which should be conditioned by the large scale spatially distributed homogeneous or heterogeneous environment with dynamic diffusion processes the multi mobile sensor systems and reconfigurable wireless networks of distributed autonomous devices which can sense or monitor physical or environmental conditions cooperatively are very sufficient. Intelligent sensors and sensor networks have an important impact in meeting environmental challenges. Agents interact (communicate, coordinate, negotiate) with each other, and with their environment. Usually, in a multi-agent system, interaction dynamics between an agent and its environment lead to emergent structure or emergent functionality [3].

There are many applications for non-stationary problems in the sense that the global optimum value and the shape of fitness function landscape (by the moving peaks) may change with time. The task for the adaptive optimization algorithm in these environments is to find optimal results quickly after the change in environment is detected.

Conceptually speaking, monitoring can be realized by continuously collecting sensory data from a distributed network of stationary or mobile Intelligent Sensor Agents deployed in the field. The architecture of such system for environment monitoring may consist of both sensors (for complex environment monitoring) and mobile Intelligent Sensor Agents, a wireless communication network [4].

Integrated sensory system is possible to treat as an information channel between the environment and the automated monitoring and mapping distributions. The development of a new range of sensor materials, effective sensors and sensory systems (networks) united in artificial intelligence techniques can achieve the necessary capabilities to provide quantitative information as well as alarm functions. Sensor networks consisted of a small number of sensor nodes that were wired to a central processing station. Sensor networks have a variety of applications. Examples include environmental monitoring – which involves physical or environmental conditions, habitat monitoring (determining the plant and animal species' population and behavior), seismic detection, military surveillance, inventory tracking, smart spaces, etc. In fact, due to the pervasive nature of micro-sensors, sensor networks have the potential to revolutionize the very way we understand and construct complex physical systems [5]. However, nowadays, the focus is more on wireless, distributed, sensing nodes. Multiple roles can be

This paper was prepared in the framework of the EU projects: SENSERA and SECURE R21

distinguished: Sensors – measure physical phenomena, sources of measurement data; Base stations – analyze and post-process data, sinks for measurement data. Actuators – perform actuation in response to received data; Processing elements – pre-processing of transmitted data.

A wireless sensor network (WSN) consists of spatially distributed autonomous sensors to monitor physical or environmental condition, and to cooperatively pass their data through the network to a main location. The WSN is built of "nodes" – from a few to several hundreds or even thousands, where each node is connected to one (or sometimes several) sensors. On the other hand, we can distinguish also two kinds of nodes: Aggregator and Device or Sensor/Actuator.

Area monitoring is a common application of WSNs. In area monitoring, the WSN is deployed over a sensor field where some phenomenon is to be monitored. When the sensors detect the event being monitored, the event is reported to one of the base stations, which then takes appropriate action. Sensor nodes can be imagined as small computers, extremely basic in terms of their interfaces and their components.

The base stations are one or more distinguished components of the WSN with much more computational, energy and communication resources. They act as a gateway between sensor nodes and the end user as they typically forward data from the WSN on to a server. The algorithmic approach to modeling, simulating and analyzing WSNs differentiates itself from the protocol approach by the fact that the idealized mathematical models used are more general and easier to analyze.

To better support high quality monitoring, we propose to enhance the sensor network with mobile swarms. A "swarm" is a group of nodes which are physically close to each other and usually share the same mobility pattern [6].

Swarm intelligence is an exciting new research field still in its infancy compared to other paradigms in artificial intelligence. Particle swarm optimization algorithms (PSO) have gained popularity in recent years. PSO is a population-based method, a variant of evolutionary algorithms with moving towards the target rather than evolution, through the search space. In PSO algorithm, the problem solution emerges from the interactions among many simple individual agents called particles [7].

The movements of the particles around in the search-space are guided by their own best known position in the search-space as well as by the entire swarm's best known position. The improvement of positions is a necessary condition to guide the movements of the swarm. The gradient of fitness or cost function, which must be optimized, is not known. The goal is to find a solution in the search-space, which would mean is the global optimum. The process is repeated and by doing so it is hoped, but not guaranteed, that a satisfactory solution will eventually be discovered.

II APPROACH TO QUANTUM INFORMATION TECHNOLOGY METHODS

Nowadays information processing is fundamentally studied with classical approaches; the latest improvements in this direction use existing explorations and no significant breakthroughs are observed. The explanation of such difficulties lies under the natural limitations to which we are already close enough. Our progress barely satisfies our needs, for we are reaching the edge of existing paradigms; consequently, we seek for novel approaches of information processing. Information processing methods based on quantum mechanical phenomenon is believed to be closer to nature, which promises to open a whole new world of opportunities. Moreover, emerging technologies use nano and sub-nano scales, where quantum mechanics comes into play, and we can't ignore its influence on the computing process and we see information processing based on quantum approaches as the future of information science [8,9].

The main actor in quantum system – the main unit for saving information- is called quantum bit or qubit. Qubit exists simultaneously in two states, and there is certain probability to measure qubit in classical state 0 or 1. After measurement, we lose the superposition and from all possible states we get just 0 or just 1. To take advantage of this property, we must operate on the qubits as long as needed and measure them only at the end because operating saves the superposition. We have restrictions on the types of operations; every operation should be reversible (intuitively it's easy to understand that quantum operations are reversible because there is no lost of information and we always can go back, reverse the process), but measurement is irreversible and all irreversible operations collapse the superposition. Furthermore, no cloning theorem tells that every particle in the universe has its own unique state. We can't fake it (nature seems to forbid making an exact copy of something). We can't hide the information a particle contains; it's somehow represented into its unique quantum state, so this can be used to detect false. It's awful if we would think about spreading information, but from the security point of view it gives novel opportunities [10].

The outline of this problem includes: a) Quantum computation (QC) – quantum bit (qubit) and entanglement; – problems in experimental realization of QC; b) Spin-based QC – nuclear spin and electron spin in semiconductors as qubits. A challenging problem is to use the reach world of correlations in quantum systems in a controllable manner to process information [11].

A quantum particle with two steady state levels can be used as a quantum bit \equiv *qubit*

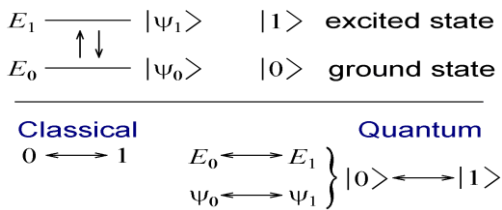


Fig.1. Examples:

- ground and excited states of an atom;
- vertical or horizontal polarization of a single photon;
- superconducting and normal state; - spin 1/2 particles in a magnetic field.

Classical bit can represent at the moment either 0 or 1. Most general qubit state is a superposition of two basic states:

$$|\psi_1\rangle = \alpha |0\rangle + \beta |1\rangle$$

$$\alpha^2 + \beta^2 = 1$$

For two bits there are four possibilities: 00, 01, 10, 11. In contrast, two qubits are in general in a state of a form

$$|\psi_2\rangle = a|00\rangle + b|01\rangle + c|10\rangle + d|11\rangle$$

$$a^2 + b^2 + c^2 + d^2 = 1$$

Qubits in this state display a degree of correlations impossible in classical physics. This phenomenon is called entanglement and is a crucial property for the success of quantum computing.

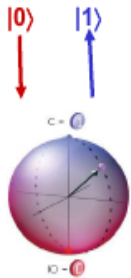


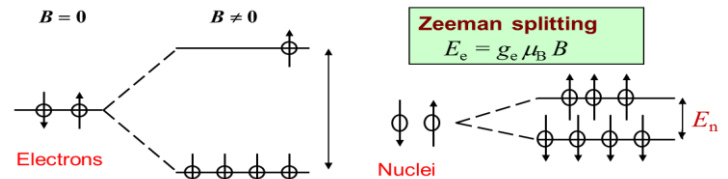
Fig.2. The general state of N qubits is specified by a 2^N - dimensional complex vector.

The main requirements for the implementation of a quantum computation are:

1. A scalable physical system with well characterized qubits: Two-level systems - spin 1/2 particle in a magnet field where one is ground and excited states of an atom, the second – superconducting and normal state.
2. Long relevant decoherence times: at least 10^4 - 10^5 times longer than the gate operation time. This is necessary for successful application of the quantum correction procedure
3. The ability to initialize qubits to a ground state, such as $|000\dots\rangle$: registers should be initialized before the start of computation.
4. A “universal” set of quantum gates: two-qubit interactions: CNOT (control not) or SWAP gates (universal quantum gates).

5. A qubit-specific measurement capability: the result of computation must be read out.

Among many suggestions for realizing the basic unit for Quantum Computation, the most exciting avenue is using spin-1/2 particles (electrons, some nuclei) embedded into a semiconductor device which allows to utilize the



tremendous resources of silicon based industry for scalable fabrication technology [12].

Fig.3. Candidate for a qubit needs phase coherence during quantum com

In the last decades our perception of the world has changed. We are more involved in distant and shared tasks, and it's natural to seek new ways to improve communication. As quantum information theory evolves and shows its great potential, why not try and take these advantages and make communication better.

Quantum communication refers to a process of transferring qubits from Point A to Point B at distance. There are a set of problems for which it's more efficient and even not possible with classical communication to solve than with quantum equivalent. The best known example is QKD, though there are quantum non-locality (entanglement), quantum teleportation, communication complexity and many more [13].

Quantum communication relies on some phenomenon like entanglement which gives plenty of opportunities, but at the same time it's very tricky. When Einstein first saw this phenomenon (EPR – Einstein-Podolsky-Rosen paradox), he said that it was incompleteness of quantum mechanics. For today, we know more about this phenomenon; still we have a lot to explore in this direction Present knowledge lets us define entanglement as a property of quantum system when two or more objects are linked together (their quantum states) and you can't refer to one without referring to the others, so if you measure one, others are determined as well. If we define communication in qubit terms, every qubit has its own channel to transmit state, but sometimes it happens so that two or more qubits are entangled and share the channel which means they communicate between each other. The result of communication is the “immediate” transmission of one of the qubit’s state to others [14].

The medium that provides the communication is unknown for today. It can be said that it is some sort of field. It is important, that this field fills the space and even more - the communication is not based on light speed, which means

either the distance does not reflect on the communication time or it reflects less.

Still, we try to use what we know about entanglement, and we came up with a strange communication scenario which we call quantum teleportation. But if you look closer, it's nothing that special.

Quantum teleportation is a great proof of entanglement's power. Quantum teleportation is a process when an object's quantum state dissolves here and reappears at a distance without ever existing at any intermediate location. Process can be executed as a three step sequence (Fig.4).

First an entangled pair of qubits are prepared and distributed, then the sender performs a so called Bell-State-Measurement (BSM) between entangled qubit and qubit to be teleported and sends measurement result to receiver via classical channel. Note that BSM provides nothing about the teleported qubit's state, but contains something about how the two are related (entangled and teleported qubits), and this information is infinitely smaller than classical description of teleported qubit's state. At the end, based on BSM results, the receiver makes result-dependent unitary rotation to his/her system to recover qubit state. So as we use the classical channel to communicate, we are limited with the speed of light and physical implementation (with linear optics) of BSM is not as efficient as needed [15].

Communication complexity refers to the number of communications required to solve a distributed task. For example if A and B points (separated in distance) have their own input x and y and want to calculate some function $f(x, y)$, what would be the minimal amount of information exchange (communication) for problem to be solved?

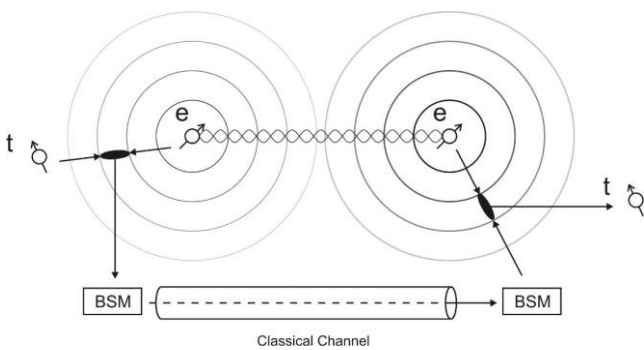


Fig. 4 Visual representation of teleportation: t – Qubit to be teleported; e – Entangled qubits; BSM – Bell-State-Measurement result.

Quantum superposition phenomenon plays a significant role in Quantum Algorithms. There are also some limitations but generally quantum algorithms are more advanced than classical ones. The most important advantage is the possibility to maintain all of the states simultaneously during the process. Theoretically it gives us the exponential power of quantum computer, but for today there are many technical and principal problems, which limit us to implement quantum algorithms [16].

III QUANTUM TECHNOLOGY METHOD OF SENSORY SYSTEMS DATA COLLECTION AND PROCESSING

Environment condition data collection process usually is managed by the different methods and devices (which determine the types of data) are used [17]. These methods might be divided as:

1. Standard measurement methods – sensors and sensory systems which are measuring the various physical and chemical parameters; data bank is inflated / significantly increased by these common and other standard data .
2. Semantic description (estimated texts) – collection of information represents the unstructured data coming from sensory systems. For their valuable use semantic analysis is necessary as well as the structuring of data knowledge taking from estimated texts and performing another range of tasks.
3. Description by multimedia sensory systems (photo and video clamping, audio recording) – the process is mainly dedicated to image recognition and may have a large range of complexity (Fig.5)

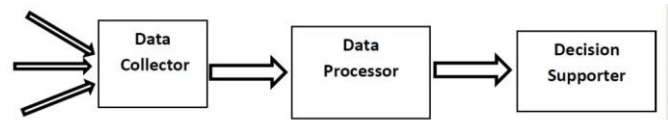


Fig.5. General scheme of the modeling process

Let us stay at the first method. In predicting the consequences of certain actions or events, various models play a crucial role. The key of simulation of the model for any purpose is determination of its tasks and targets.

Data collection is one of the most important parts of the model successful functioning. Quantum approach for optimization of the use of databases can give to us at least two positive effects: 1. The compact representation of the database; 2. The possibilities to reduce the processing time.

It is well known in computer science the representation of the three main types of data: real-valued, integer or Boolean. These types reset all data that belongs to the type of class one (Classes of data type): primitive data types (machine data types, Boolean type, numeric types); composite types (enumerations, string and text types); other types (pointers and references, functional data types, abstract data types; utility types).

Primitive or composite data types are used in the models describing the data of disaster types. This depends on two main factors: 1. Data source and 2. Model representation of type Boolean (logical), an integrated, causal, and the others.

Any type and scale of the disaster we can imagine, as it is a big system. One characteristic of this kind of systems that are used to describe the condition of many and varied attributes and, in addition, may consist of many smaller ones define the subsystems. For example, if the object of our study is earthquake, earthquake could lead to flood or landslides [18]. Therefore, we have at least two different and mutually dependent systems, which at the same time could be described by a separate model.

The parameters of each of this kind system are divided into descriptive qualitative parameters (include only content definitions), quantitative parameters (include only discrete or continuous quantitative parameters) and mixed parameters (include quantitative and qualitative parameters all together). In the real situation when disaster is mixture of different systems of parameters it is very important to use a method where all types of data are in the form of inference and therefore there are no information loss and all these can be used in a single model. Let us represent a generalized notion of the catastrophe. To observe the different disasters jointly because of the reason of their high individuality is not always possible [19].

Let us admit that we have S a big system, and its describing parameters are:

$$x_i \in \{P\} \cup \{C\} \cup \{A\}, i = 1, \dots, N,$$

the abundance of (where N is the description of the parameters of the points, {P} - Primitive data type, {C} - Composite data type, {A} - Abstract data type) model is essential to reducing the effectiveness of the same type. Type depends of the chosen model. In our case, all parameters are reset to the quantum dimension of the quantum value.

We can perform the transformation process in two stages: 1. Unification of logical presentation of data from the census; 2. the quantum representation of parameters.

In the first step, for each parameter there is a discrete set of values, which contains much of numeric values [20], contextually described in non-overlap range. The question is; how many different values can be fixed when we describe the S system, which was adopted by the International grading system describing or defining the level of threats. This number can be different for each parameter. As a result, we get the allowable values for each parameter draws x_i domain:

$$x_i \in \{x_i^1, x_i^2, \dots, x_i^{n_i}\},$$

where n_i is equal to x_i a number of different meanings .

In case that we have the S system description in the generalized form we could say that the description of the observed x_i is the main option or not. In some cases,

different it is important first to analyze the existence of zero in the option. In this case the quantum x_i performance is used.

Assume that x_i is a quantum imagination of the system, which can be represented as $|x_i\rangle$ and the state of a quantum system $|x_i\rangle$ is a vector in a complex vector space (Fig.6). If the state of a quantum system $|x_i\rangle$ is a vector in a complex vector space and the set of vectors $\{|n_i\rangle\}$, $n_i = 0, \dots, N - 1$ (where N may be ∞) has an orthonormal basis, for this space we can always express

$$|x_i\rangle = \sum_n c_n^i |n_i\rangle$$

for some complex coefficients c_n^i , where $\sum_n |c_n^i|^2 = 1$.

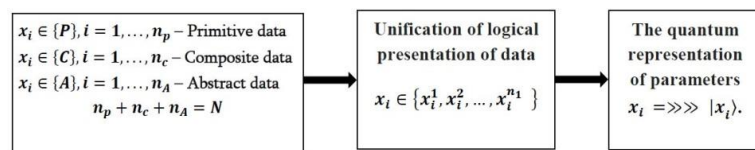


Fig.5. The data transformation process.

According to the parameters of each presentation of each $x_i x_i^j, j = 1, \dots, n_i$ meaning we can write: c_j^i ratio. This form will be modified according to our data base for further processing [21].

In the point of view of data processing, we see two approaches: 1. Quantum information processing classical quantum algorithm (mainly meant to search), and 2. Data processing system S is used to describe and convey it by the quantum concepts.

Grover's algorithm for quantum calculations is one of the most important tool, which helps to describe not well defined $N=2^n$ elements in the database (a database handle disasters) of a particular element search. This algorithm makes possible to compute the many unsolved problem of classical calculations [22]. Using classical methods in the theory of probability, we can say that for any m element inspection the probability that a request for records with equal m/N . It is clear that the database needs to be $O(N) = 2^n$ for the necessary elements to look for. Using the Grover's algorithm the necessary number of requests (steps) is $O(\sqrt{N}) = O(2^{\frac{n}{2}})$

In our case, we may give to the task such formalization: S is used to describe the system of $N=2^n$ From the each of the values S_1, S_2, \dots, S_n , there is a unique situation, which

satisfies the condition: $f(s_u)=1$ is only one element $s \in A$, and $f(s)=0$ for all other elements. We can make such a formulation of the problem, because we have already reduced our options to quantum face.

As mentioned above, we use the S parameters for description of the $x_i, i = 1, \dots, N$. In concrete case presentation these parameters reducing to the quantum face. Suppose we have a different system in $K S$ Description. Each description we can write as quantum implications: Where

$$|\widetilde{x_1}\rangle \& |\widetilde{x_2}\rangle \& \dots \& |\widetilde{x_N}\rangle$$

$$|\widetilde{x_i}\rangle = \begin{cases} |x_i\rangle & \text{description of the } S \text{ system;} \\ |\overline{x_i}\rangle & \text{no description of the } S \text{ system.} \end{cases}$$

Therefore we have K implicants. Write a realization in dysfunctional normal form:

$$\bigvee |\widetilde{x_1}\rangle \& |\widetilde{x_2}\rangle \& \dots \& |\widetilde{x_N}\rangle$$

If we minimize this form following the method which it is shown in [23], as result we will receive S system describing quantum concept in a generalized form. This description is compact and contains only those parameters that are most important to a particular kind of system evaluation. Its use will enable to evaluate the system not only by quantitative parameters, but options of all of them.

Quantum algorithms (especially Shor's) prove that quantum approaches are more flexible than classical in complex environments like the sensory network (when process goes exponentially). So we can say that our tool-set of information processing (brain) must be minimum quantum, as far as we know, because Devo (development evolution) have chosen us as leading creatures.

Shor's algorithm have suppressed any hope that encryption base on discrete logarithm (factoring large numbers) can be resistant against quantum computing, so we have to replace asymmetric encryption algorithms with novel quantum approaches. Though it was proved that symmetric algorithms perform quite well in quantum environments, similar approaches do not give effective use in quantum asymmetric world. Instead of pure asymmetric key distribution there are some thoughts about quantum asymmetric cryptography using entangled key pairs. This approach effectively uses the physical security of channel, so to estimate private key with high probability eavesdropper needs large amount of public keys. The disadvantage of this approach is the need of trusted issuer of keys, who generates private and public key pair and sends it to authenticated users securely [24].

Suppose eavesdropper reads quantum state in channel without collapsing superposition. QKD uses quantum channel to negotiate the key, after that encrypted data is sent

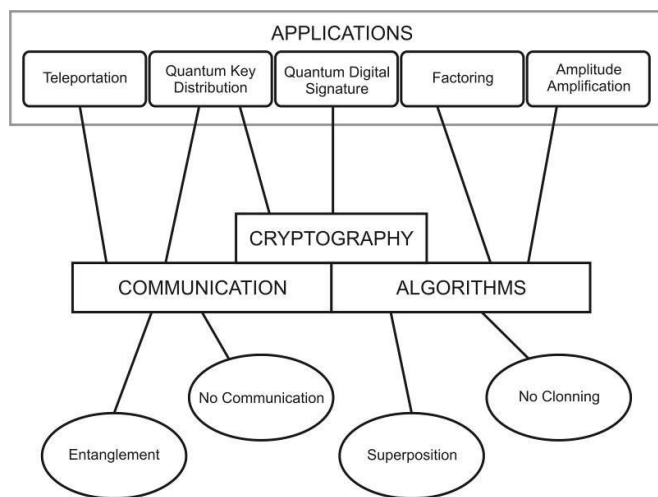
via the classical public channel. The security of this method will become vulnerable if quantum reading without losing superposition is possible. The next advancement of attack on quantum channel is to get the information with about the same probability as the receiver. One of the known attack modes includes man-in-the-middle attack at any point in QKD. The reliability of QKD is based on encoding the information in non-orthogonal states. Quantum indeterminacy means that these states cannot be measured without disturbing the original state. However, if we read superposition without collapsing it, then we would be in the same state as actor A, thus if man-in-the-middle attack continues, we can get the key generated using QKD.

If we assume that reading qubit is possible with certain (high) probability and the eavesdropper can easily access information sent by quantum channel, the need of different approach arises. One protection against such attack is the use of entangled qubit pairs as key. The sender generates entangled registers (sequence of qubits) and sends one of them via quantum channel. The eavesdropper can successfully read this information but after the receiver receives it, sender operates the entangled register on his/her side and in that way sets a key, which is also accessible to the receiver but is unknown for the eavesdropper. After that, encryption and decryption happens with entangled keys. If we assume that entanglement could not be intruded by a third-party, then this scheme is unconditionally secure.

Let's take n -bit qubit and apply some algorithm on it. After performing the main algorithm quantum register is in superposition of all possible states. The goal is to find the solution, that is one of the $N = 2^n$ states. Let's take the simple model to describe amplitude amplification. The best example should be uniform superposition that contains every possible state with the same amplitudes (coefficients) and the sum of the squares of amplitudes is 1. The solution we are interested in is one of them, but if we measure the register, the probability of getting solution is $1/N$. We could try again and again, but which one would be a solution is unknown, so we can't effectively find the answer without changing the coefficients. This is the case where amplitude amplification comes into play. We use the "oracle", which changes only the solution's amplitude. Applying oracle on qubit result in the change only in the solution's amplitude; specifically, it gets a minus sign, so the probabilities (square of amplitudes) remain unchanged. After that the amplitudes are inverted against the mean of amplitudes; consequently the solution's amplitude is raised and the other amplitudes are lowered. The reason this happens is that only the solution has negative amplitude which is less than the mean, and all remaining amplitudes are more than the mean. The above described steps make one iteration of Grover's algorithm [24]. When we apply Grover's algorithm, we change the amplitudes iteratively, so that on every iteration the amplitude of solution is changed. This process is periodic. Not every iteration will raise the amplitude. To be more specific, within

$$r \approx \frac{\pi}{4} \sqrt{N}$$

steps the amplitude is increased, but on r+1 step, amplitude begins to decrease. This means that we need r iterations to get maximum possible amplitude effectively. The complexity is sub-linear and is $O(N^{1/2})$ which is better to simply repeat the main algorithm several times and analyze measured results to "guess" which result is solution. There is one interesting detail about the oracle. Oracle is represented as a matrix which contains "1"-s on the diagonal except one element which is related to the solution and is "-1" and all other elements are "0". In the real world, we don't know where that "-1" is, because if we knew, we also would know the solution itself. The oracle hides the solution in itself; we just can use it to increase the probability of measurement.



Finally, it's worth mentioning, that finding the quantum solution is more effective if the states are unstructured and unsorted because in sorted cases there are no significant differences between classical and quantum algorithms. So, Grover's algorithm is used for searching the solution that is already mixed in the superposition of quantum register; more generally it's useful for searching one particular element in an unsorted, unstructured set. This algorithm is also expanded to search for multiple solutions in the superposition, but this is beyond our scope [25].

Applications we have reviewed, we think are ready for mainstream after implementing the quantum computer. The approaches that lie under these applications use quantum properties (Fig.5).

Fig. 5 Quantum Approaches and its applications.

Cryptography has always been an important part of information theory. A lot has been done in classical

cryptography. Two main types of algorithms are known – symmetric and asymmetric. Symmetric algorithms are widely used for securing communication between two parties (e.g. A and B) and asymmetric are used for digital signatures. Both of them have their advantages and disadvantages, but we will discuss classical cryptography only in order to explain quantum cryptography. Security in classical cryptography was based on the exponential number of computational calculations, which was not achievable in real time, needed to decrypt encrypted message. As we have seen, quantum computing offers us exponentially more computing power than classical analogy. Due to that, many algorithms which were thought to need years to break the key, need no more than minutes with the help of quantum computation. But that will be done after the quantum computer is built. Until quantum cryptography can help these problems, it solves classical cryptography's weak issues. Quantum cryptography is mainly known for quantum key distribution (QKD) which solves the weakest point of classical symmetric algorithms- key distribution [26]. Despite this fact, integration in classical cryptography is essential because QKD only generates and distributes keys over two parties which then can use this key with any classical encryption algorithms.

The idea of QKD algorithm is the following:

We begin with the first stage, the transmission of the photons, which is the physical representation of qubits, from A to B. Afterward the communication switches to the public channel. There, the first phase is the shifting phase, where A and B negotiate which bits are used and which bits are discarded. To avoid a man-in-the-middle attack by C, this message exchange must be authenticated. After agreeing on the bits and being sure that C has not modified messages by using an authentication scheme, A and B go on to the reconciliation phase or error correction phase. Due to the fact that quantum channel is not a noiseless channel, A and B do not share the same identical string. There is a small portion of errors in B's string, which are corrected in this phase. Again, C has the possibility to modify messages during this phase to his/her interest. Therefore, A and B must authenticate this phase. Passing reconciliation, A and B share a string, which is identical with a very high probability. But this string cannot be used as a key yet. C's information about the string must be considered.

As we can see QKD is limited in distance, because in the first part we send qubits via quantum channel. Due to this fact, no cloning theorem and no repeaters can be used as in classical communication. This type of algorithm is unconditionally secure if several simple conditions are met:

1. Eavesdropper cannot access A's and B's encoding and decoding devices
2. The random number generators used by A and B must be trusted and truly random (for example a Quantum random number generator)

3. The classical communication channel must be authenticated using an unconditionally secure authentication scheme

Despite this fact QKD has been broken, but not because of the algorithm but due to the non ideal behavior of the present-day quantum cryptographic hardware [27].

As for asymmetric algorithms some theoretical advancement is present. Nothing has been done in practice because asymmetric quantum algorithms require quantum technology beyond today's advancements.

Quantum digital signature algorithm is already available.

Quantum digital signature shares a lot with classical analogy. Requirements for good and usable signature schemes for classical and as well as for quantum are underlined:

1. The scheme has to provide security against tampering by: - The sender after the message was signed; -The receiver; - A third party
2. Creating a signed message has to be easy
3. Every recipient has to get the same answer, when testing the message for validity

Differences between classical and quantum signatures are based on quantum information nature.

IV CONCLUSION

This work was motivated by the idea of developing the high effective sensory systems monitoring of environmental pollution, particularly in nuclear power engineering, which can be realized by continuously collecting sensory data from a wireless mobile sensor network deployed in the field [28]. The relevance of problems is particularly pointed out by the environmental dynamism of the shape of fitness function landscape, which consists of a number of peaks of changing width and height and in diffuse processes. We have discussed the quantum algorithms as effective tools for the adaptive control of the mobile sensory system .

We also discussed the quantum approach of sensory data collection and processing using some quantum information technology methods and tools.

Present knowledge lets us define entanglement as a property of quantum system when two or more objects are linked together (their quantum states) and you can't refer to one without referring to the others, so if you measure one, others are determined as well. More than that, no matter how far they are (physical separation), measurement occurs instantly, faster than the speed of light. It's like an instant communication which would be great, so that we could reduce the dependency on distance, but unfortunately lack of knowledge does not allow us to realize its potential. Inside the quantum world, entanglement is some sort of communication because the separated states depend on each other or have a connection.

Taking into account the quantum and multi parametrical nature of information for its clear and precise modeling it is

possible and effective to combine two methods, where one is based on quantum numbers usage for performing of different parametrical values (transferring logical numbers to quantum numbers), and second - to creation of the data base (quantum data base) which should be presented as its quantum model. These approaches jointly with quantum search algorithms and quantum query algorithms are opening the new ways for creation of novel technologies for modeling and creation of novel high effective sensory systems and networks.

REFERENCES

- [1] M.Roza, J.Voogd, D.Sebalji, "The Generic Methodology for Verification and Validation to support acceptance of models, simulations and data". The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology October 2013 10: 347-365
- [2] P. Kervalishvili, B. Meparishvili. "Synergy, entropy and sustainable development". Georgia Chemical Journal, vol.10, N 4, 2010, 169-173.
- [3] P. Kervalishvili, B. Meparishvili, G. Janelidze. "Adaptive control of mobile information system". NATO Science series, IOS press, v.93, 2012, 100-108.
- [4] Luís M. L. Oliveira and Joel J. P. C. Rodrigues. "Wireless Sensor Networks: A Survey on Environmental Monitoring." Journal of Communications. Vol. 6, no. 2, April 2011.
- [5] Wilson, D.M. S. Hoyt, J. Janata, K. Booksh, and L. Obando, "Chemical Sensors for Portable, Handheld Field Instruments." *IEEE Sensors J.*, Vol. 1, No. 4, 2001, 256-274
- [6] K. E. Parsopoulos, M. N. Vrahatis. "Particle Swarm Optimization and intelligence: Advances and Applications." Published in the United States of America by Information Science Reference. Hershey, New York ISBN 978-1-61520-666-7, 2009.
- [7] Y. H. Shi, and R.C. Eberhart, "Empirical Study of Particle Swarm Optimisation," Proceedings of the Congress on Evolutionary Computation, (Washington D.C. USA), IEEE Service Centre, Piscataway, NJ, 1995, 1945-1949.
- [8] P. Kervalishvili. "Philosophy of quantum information science". NATO Science series, IOS press, v.93, 2012, 55-73.
- [9] P. Kervalishvili. "Quantum Information Science: Some Novel Views." In Book *Computer Science Technology* an Applications. Nova Science Publishers, Boston, USA, ISBN: 978-161324-870-6, 2011.
- [10] Alexander Holevo. "Quantum Informatics." Science World (Scientific American - Rus.), No7. 2008.
- [11] Paata J. Kervalishvili. "Development of Quantum Information Technology based on nuclear spin qubits". The Eighth Japanese-Mediterranean Workshop on Applied Electromagnetic Engineering for Magnetic, Superconducting, Multifunctional and Nanomaterials

(JAPMED'8), book of abstracts NCSR Demokritos, 24-26 of June, 2013, Athens, Greece.

[12] B. Kane, "Si-based nuclear-spin quantum computer", *Nature* 393, 1998, 133.

[13] A. Liang, V. Scarani., J. G. Rarity, J. L. O'brien, 2010. Reference Frame "Independent Quantum Key Distribution". Centre of Quantum Photonics, University of Bristol, U.K. Centre of Quantum Technologies and Department of Physics, National University of Singapore, Singapore, arXiv:1003.1050v1

[14] C. H. Bennett. "Notes on the History of Reversible Computation". *IBM J. Res. Dev.* Vol. 44. No ½, Jan.-March. 2000.

[15] Nicolas Gisin, and Rob Thew, "*Quantum Communication*", Group of Applied Physics, University of Geneva, Switzerland, arXiv:quant-ph/0703255v1 2008.

[16] D. A. Meyer. "Physical Quantum Algorithms". UCSD preprint, 2001.

[17] D. Estrin, L. Girod, G. Pottie, M. Srivastava, "Instrumenting the world with wireless sensor networks." In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP 2001)*, Salt Lake City, Utah, May 2001.

[18] D. Pfeifer, J.Valvano, A.Gerstlauer. "SimConnect and SimTalk for distributed cyber-physical system simulation". *SIMULATION*, October 2013 89: 1254-1271.

[19] Paata Kervalishvili, Manana Khachidze. "Quantum Information Technology and Modeling of Disasters – A Prospective View". Presentation at the NATO Advance Research Workshop – ARW, Improving Disaster Resilience and Mitigation - New Means and Tools, Trends Jassy Romania, November,6-8 2013.

[20] Kazuo Iwama. "Quantum Search Algorithms for Database Query Processing". ERATO Quantum Computation and Information Project. September 6-8, 2001, Tokyo, Japan. qci.is.s.u-tokyo.ac.jp/qci/eqis/Iwama.ps

[21] Sudip Roy, Lucja Kot, Christoph Koch. "Quantum Databases". *CIDR*, 2013.

[22] Archuadze M., Besiashvili G., Khachidze M. and Kervalishvili P. "Knowledge Engineering: Quantum Approach", published in *Philosophy and Synergy of information: Sustainability and Security*, Publication is supported by: The NATO Science for Peace and Security programme Sub-Series E: Human and Societal Dynamic-, , vol.93 ISSN 1874-6268, 2012 pp.175-185.

[23] M.Khachidze, M.Archuadze ,G.Besiashvili "The Method of Concept Formation for Semantic Search". 7th International Conference on Application of Information and Communication Technologies., Baku, Azerbaijan, 23-25 October 2013.

[24] P. W Shor. "Algorithms for Quantum Computation: Discrete Logarithms and Factoring." in S. Goldwasser, ed., *Proceedings of the 35th Symposium on Foundations of Computer Science*, Santa Fe, NM, 20{22 November, Los Alamitos, CA: IEEE Computer Society Press 124{134, 1994.

[25] L. K. Grover, "A Fast Quantummechanical Algorithm for Database Search". in *Proceedings of the 28th Annual ACM Symposium on the Theory of Computing*, Philadelphia, PA, (New York: ACM 1996) 212{219, 1996.

[26] Fei Gao, Qiao-Yan Wen, Su-Juan Qin, Fu-Chen Zhu. "Quantum Asymmetric Cryptography with Symmetric keys", arXiv:0809.3408v2,2008

[27] L. Lydersen, J. Skaar. 2010. "Security of Quantum Key Distribution with Bit and Basis Dependent Detector Flaw.]" Norwegian University of Science and Technology, Trondheim, Norway, University Graduate Center, Kjeller, Norway, arXiv:0807.0767v4

[28] Paata J. Kervalishvili and Tamara M. Berberashvili. "Quantum Effects Based Nanosensory Systems". *Black Sea Energy Resource Development and Hydrogen Energy Problems. NATO Science for Peace and Security Series-C; Environmental Security*. Springer. 2013, pp.359-372.

The Effect of Using Channel Equalizer in The SDR Modem

Dr. Sabah N. Hussein

Department of computer engineering techniques
College of Electrical and Electronic Engineering Techniques
Baghdad, Iraq
drsabah2004@yahoo.com

Raghad S. Majeed

Department of computer engineering techniques
College of Electrical and Electronic Engineering Techniques
Baghdad, Iraq
raghad_eng89@yahoo.com

Abstract— The channel equalization is a technique allowing to remove the inter-symbol interference (ISI) in the SDR receiver caused by the radio-mobile channel. In this paper, present a design and implementation of modem with 16-QAM modulation, convolution, interleaved circuit, differential coding and adaptive equalizer using LMS algorithm based of SDR, using MATLAB system generator model. The results show that LMS equalizer gives a good estimation on the proposed channel model. The hardware implementation is done in FPGA board kit, which has shown a promising foundation for developing coding, modulating and other circuit of modems circuits.

Keywords— FPGA; LMS equalizer; ISI; SDR modem; 16-QAM.

I. INTRODUCTION

Software defined radios (SDR) are highly configurable hardware platforms that provide the technology for realizing the rapidly expanding future generation digital wireless communication infrastructure [1]. In many practical communication systems, data is transmitted over a channel with inter symbol interference (ISI). To reduce ISI in proposed SDR system, two methods are suggested that are convolution code and least mean square error algorithms [2]. Convolution code known also as error correcting code added redundant bits to the information transmitted bits to allow the receiver to detect and correct a limited number of errors occurring anywhere in the transmitter signal. Different studies that are combined convolution code in SDR system like in [3]. Equalizer gives the inverse of the channel to the received signal and combination of channel and equalizer gives a flat frequency response and linear phase [4]. The LMS algorithm is a type of the adaptive filter used to discover the filter coefficients in the adaptive manner that is used to model the inverse channel and overcome ISI problem. There are Different studies that are combined LMS equalizer in SDR system like in [5]. The reasons for choice LMS algorithm are simplicity; low computational complexity and better performance in may run environment [6]. The paper is organized as following: Section II provides an overview of 16-qam modulation/de-modulation, Section III provides an overview of the convolutional codes. Section IV the interleaver and de-interleaver, Section V provides an overview

of the differential coding, Section VI provides an overview of the LMS equalizer algorithms, Section VII provides the block diagram of the 16-QAM SDR system, Section VIII provides the simulation results, Section IX conclusion.

II. THE 16-QAM MODULATION/DEMODULATION

QAM is one of the widely used modulation techniques because of its efficiency in power and bandwidth. The constellation diagram of 16-QAM in contain 16 different symbols each having a different real and imaginary component, Each constellation point can represent four bits, with two bits on the I axis and two on the Q axis . The 4 bits Gray coded that represent one point in the constellation diagram can be regarded also as two of two bits words on I-axis, and Q-axis respectively as shown in Fig (1) [7].

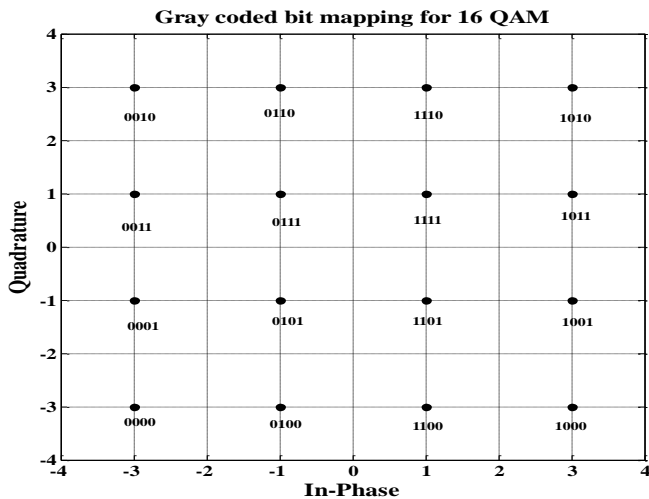


Fig. 1. The 16-QAM constellation diagram with Gray code input mapping.

The received complex coded sequence is;

$$y = x + n \quad (1)$$

x is the data complex sequence in the form of;

$$\alpha_{16 QAM} = \left\{ \begin{matrix} \mp 1 + \mp 1j, \mp 1 + \mp 3j \\ \mp 3 + \mp 3j, \mp 3 + \mp 1j \end{matrix} \right\} \quad (2)$$

In a demodulation the Maximum A posteriori Probability (MAP) method has been used as soft bit detection for 16-QAM. This method usually maximizes the probability that assume the bit b_m was transmitted given y received;

$$P(b_m/y) = \frac{P(y/b_m)P(b_m)}{P(y)} \quad (3)$$

The detail description of soft bit detection is summarizing in [8].

The soft bit for bit b_0 is;

$$sb(b_0) = \begin{cases} 2(y_r + 1), & y_r < -2 \\ y_r, & -2 \leq y_r < 2 \\ 2(y_r - 1), & y_r > 2 \end{cases} \quad (4)$$

And for bit b_1 is;

$$sb(b_1) = \begin{cases} y_r + 2, & y_r \leq 0 \\ -y_r + 2, & y_r > 0 \end{cases} \quad (5)$$

The soft bit for b_2 is similar to soft bit for b_0 except the resolutions are based on the imaginary component and b_3 is similar to b_1 but also are based on the imaginary component.

III. THE CONVOLUTIONAL CODES

Coding is a technique where redundancy is added to the original bit sequence to increase the reliability of the

communication [9]. The following parameters that can be used for convolution code are summarized as: code rate is 1/2, constraint length $K=3$, and Generator polynomial is $G= [7 \ 6]$. Fig.2 shows convolution encoder with given generator polynomial.

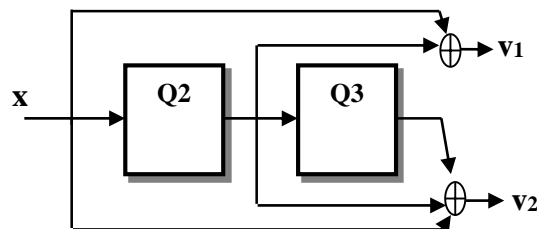
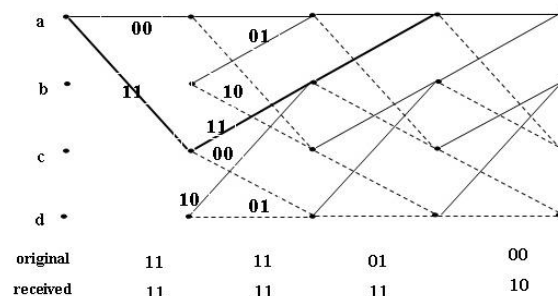


Fig. 2. The convolution code with the generator polynomial [7 6].

A. Viterbi Decoding of Convolution Codes

Viterbi Decoder algorithm is used to recover the information sequence at the receiver side. Hard decision decoder with Hamming distance measure has been used in this paper. The receiver employs a trellis based maximum likelihood Viterbi decoder which decodes the input bits to obtain the information bits. The trellis length is chosen to be 5 times the constraint length [10]. Fig (3) shows an example of



Viterbi algorithm and how the information is recovered.

Fig. 3. The Viterbi algorithm example.

IV. THE INTERLEAVER / DE-INTERLEAVER

A method for making data recovering more efficient by rearranging or renumbering. The interleaver/de-interleaver is used to reduce the effects of long burst errors. Simple random error correction code by rearranges the elements of its input vector using a random permutation [11].

V. THE DIFFERENTIAL CODING

Bit streams through transmitter can be un-intentionally inverted. Most signal processing circuits cannot know if the stream bit is inverted or not. Differential Encoding is used to protect against this possibility [12]. It is one of the simplest forms of error protection coding done on a baseband sequence prior to modulation. Supposing that x_i is a bit intended for transmission, and y_i is a bit actually transmitted (differentially encoded) [12], if

$$y_i = y_{i-1} + x_i \quad (6)$$

Is transmitted,
 then on the
 decoding side,

$$x_i = y_{i-1} + y_i \quad (7)$$

can be

reconstructed,
 where + is
 modulo-2
 addition.

VI. THE LMS EQUALIZER

The structure of the adaptive

filter is shown in Fig. (4). Least mean squares (LMS) algorithms are a class of adaptive filter used to mimic a desired filter by finding the filter coefficients that is related to producing the least mean squares of the error signal [14]. The LMS algorithm performs the following operations to update the coefficients of the FIR filter:

1. Calculate the output signal Y(K) of the FIR filter. The output of the filter represents an estimate of the desired response. Y(K) is the calculated as the convolution of the weight vector and the input vector:

$$Y(k) = \sum_{n=0}^{L-1} W_n(k)x(k-n) = W^T(k)x(k) \quad (8)$$

2. The error signal e(k), is estimation error defined as the difference between the estimated response and the desired response.

$$e(k) = d(k) - y(k) \quad (9)$$

3. The error signal and the input signal are applied to the weight update algorithm to updates the filter coefficients [13].

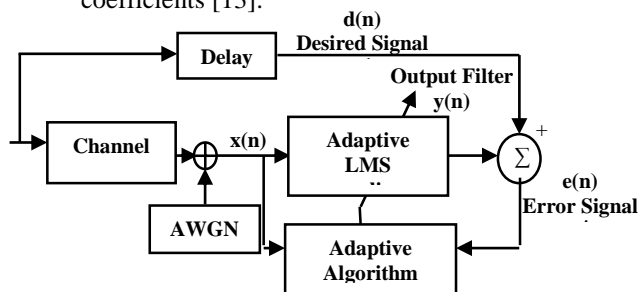


Fig. 4. The linear adaptive LMS equalizer.

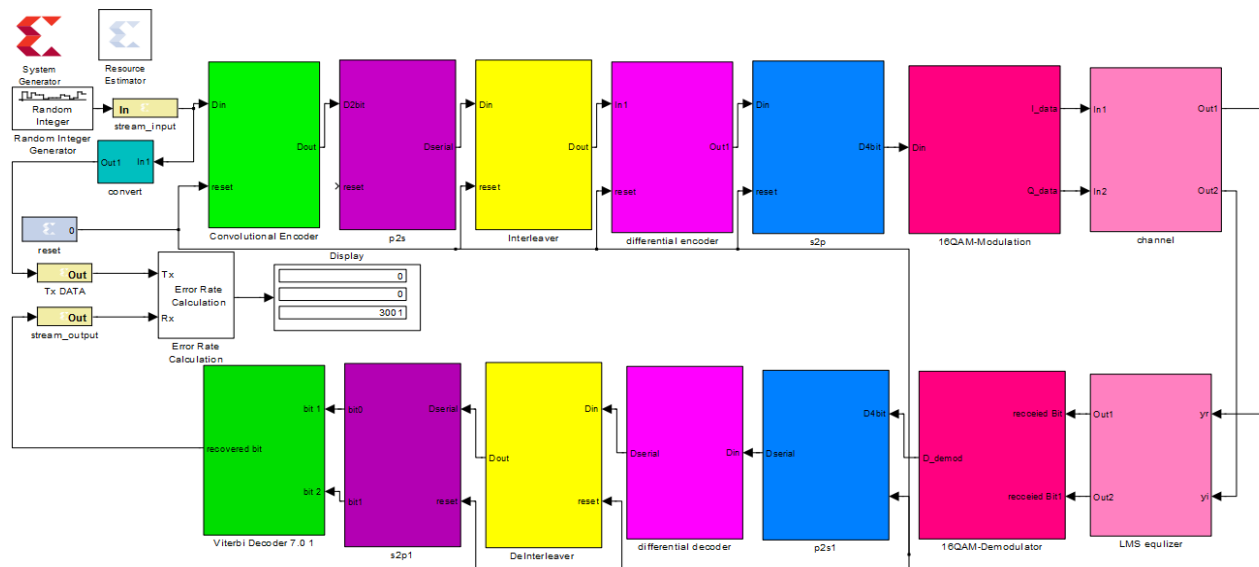


Fig. 5. The block diagram of the Baseband SDR Modem.

VII. THE 16-QAM SDR BLOCK DIAGRAM

The design of blocks of Fig. 5 can be explain based on system generator of Xilinx that work under the environment of MATLAB Simulink for FPGA design. The Past experience with Xilinx FPGA or Hardware properties Languages (HDLs) is not needed when using System Generator. All of the downstream FPGA implementation procedures including synthesis, position automatically executed to generate an FPGA programming.

A. The main blocks of transmitter section are:

- Random Binary Signal Generation

The Random integer block in MATLAB Simulink is used as a stream binary signal with 2 M-ary number .

- Convolution Encoder

The convolution encoder Xilinx IP core of system generator has been used that have native rate of 1/2, a constraint length equal to 3, and generator polynomials codes G1=110 and G2=111. Fig. 6 shows convolution encoder system [15].

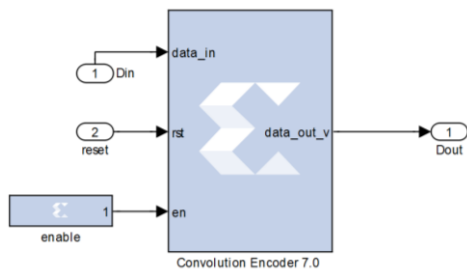


Fig. 6. The convolution Encoder Xilinx IP code.

- Parallel to Serial

The parallel to serial conversion has been done using a special converter serial to parallel available in system generator whose block layout is shown in Fig. 7. In this Block, the parallel two bits output data of convolution encode are converted to serial streams bits.

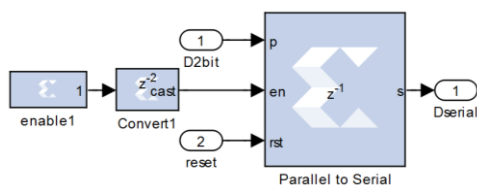


Fig. 7. The parallel to serial converter.

- Interleaver encoder

The main idea of the Random Interleaver is rearranges the elements of its input vector using a random permutation. The output of convolution code is firstly converted into serial bits and then passing through a random interleaver that is a one-to-one permutation map according to random labeling sequence after the data converted into vector mode.

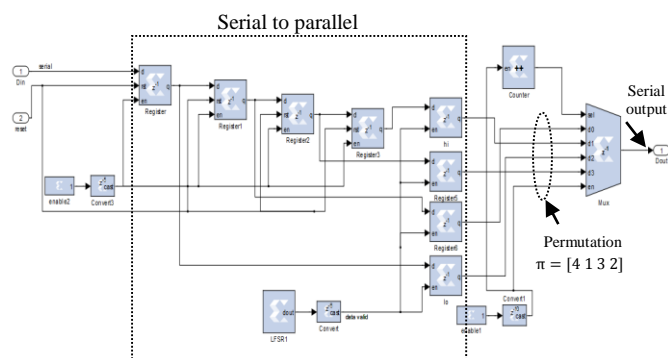
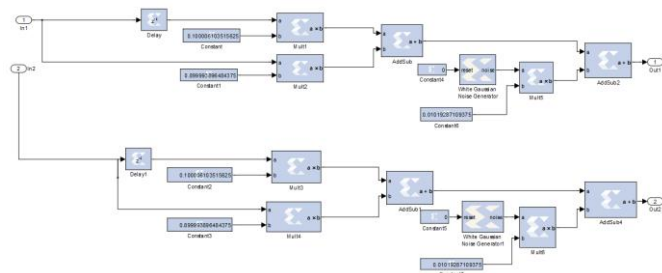


Fig. 8. The interleaver circuit.

- Differential Encoder



This encoder has been done using one delay with the logical (exclusive OR) components as shown in Fig.9.

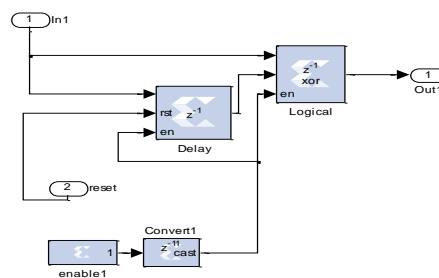


Fig. 9. The differential encoder circuit.

- Serial to Parallel converter

This block has been used to convert the Din serial bits to D4 bit parallel 4 bits of data in the form d (3:0) using shift registers as shown in Fig. 10.

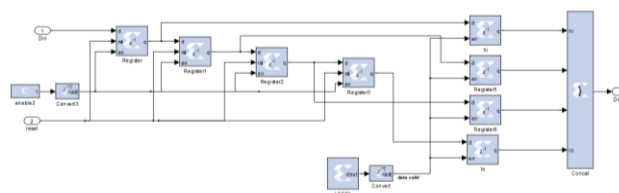


Fig. 10. The serial to parallel converter block.

- 16-QAM Mapping

Each parallel four bits generated from the serial to parallel section are mapped using the 16-QAM constellation. The four

coding values (± 1 to ± 3) are stored in a ROM memory block. The block diagram of 16-QAM mapping is demonstrated in Fig. 11 [6].

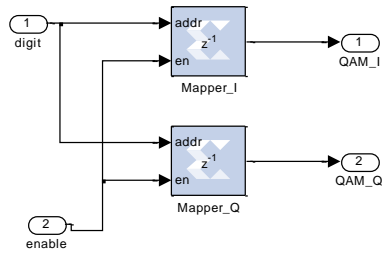
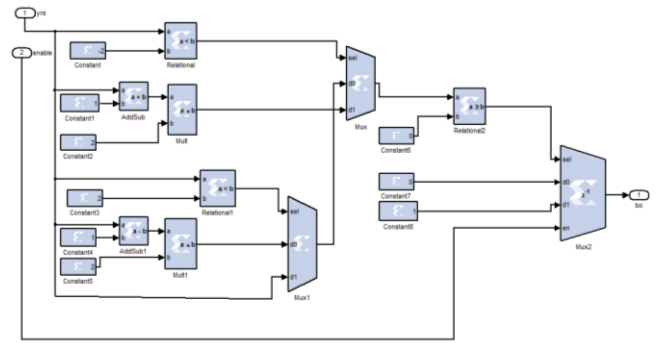


Fig. 11. The 16 QAM Mapping.



B. The main blocks of channel model:

To modeling the multipath channel two taps of coefficients of 0.1 and 0.9 respectively have been used with AWGN. The signal to noise ratio has been set to (SNR= 40 dB).

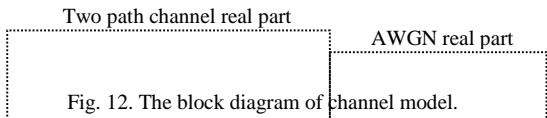


Fig. 12. The block diagram of channel model.

C. The main blocks of equalizer filter:

The LMS adaptive equalizer algorithm is shown in Fig. 13, in this Fig. x represent the desired signal that used for training, and y complex represent the actual samples. The stream sample pass through shift register to select five samples as shown in Fig. 14; the five samples are entered in parallel to the LMS filter to update weight. To update weight firstly multiplied YK with Mu by complex multiplication then adding to the older weight to produce the new weight as shown in Fig. 15. output of LMS filter that is feedback to subtract with the next complex desired signal to produce the error.

Fig. 13. The block diagram of the equalizer filter.

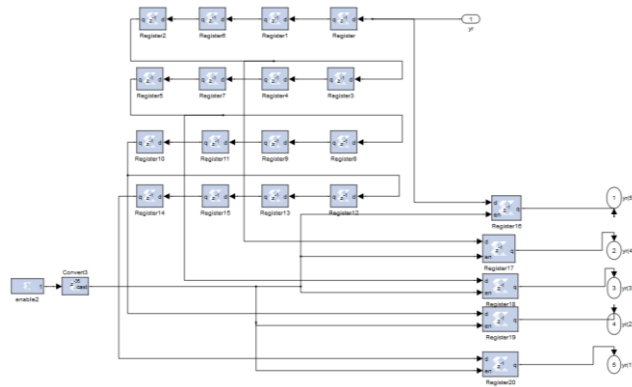


Fig. 14. The block diagram of the shift register

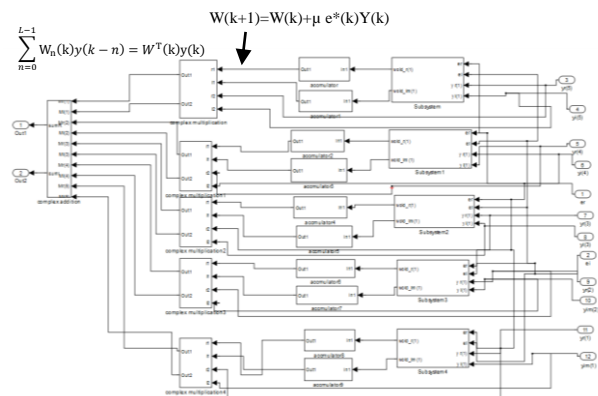
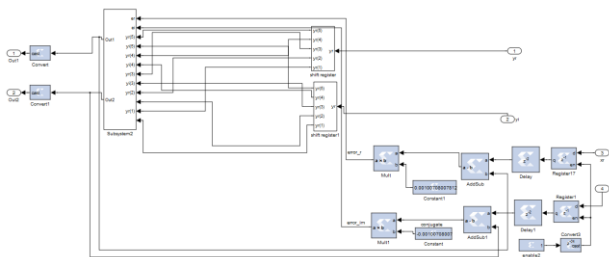


Fig. 15. The block diagram of LMS filter and update weight block.

D. The main blocks of receiver section :

- 16 QAM De-Mapping

The De-Mapping has been performed by assigning the received I-Q signals location to the nearest point in the I-Q constellation using soft bit algorithm. Figs. 16 and 17 shows the soft bit decision circuit for b_0 and b_1 respectively. Soft bit decision of b_2 and b_3 has has been built in the same manner as b_0 and b_1 respectively [7].

Fig. 16. The Soft bit decision of 16-QAM de-mapping for b_0 .

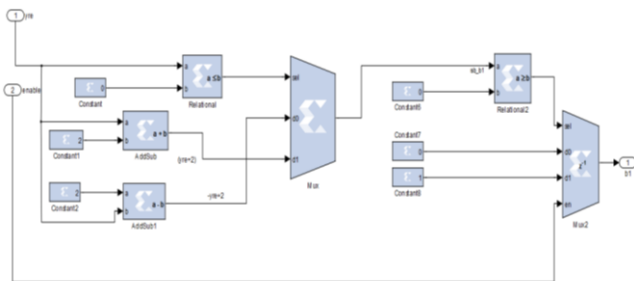


Fig. 17. Soft bit decision of 16-QAM de-mapping for b_1 .

- Parallel to Serial Conversion

The parallel four bits output from 16 QAM De-mapping has been converted to serial stream bits using parallel to serial converter as shown in Fig. 18.

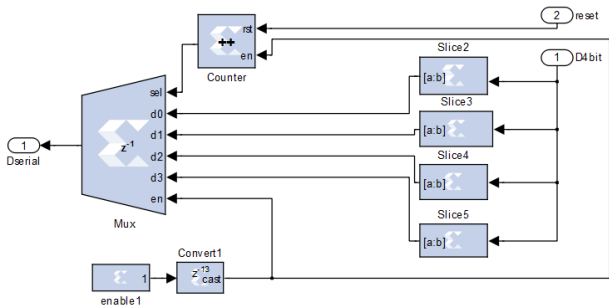


Fig. 18 The parallel 4 bits to serial converter.

- Differential Decoder

Fig. 19 shows the differential decoder circuit. Using one bit delay block, and one block exclusive OR components [3].

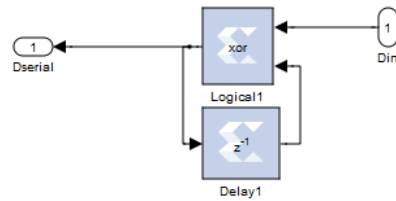


Fig. 19. The differential decoder circuit.

- Random De-interleaver

The output of the differential decoder is stream bit and each four bits serial that convert to the parallel with re rang the sequence for each bit it's the inverse for the interleaver encoder and the output is serial bit.

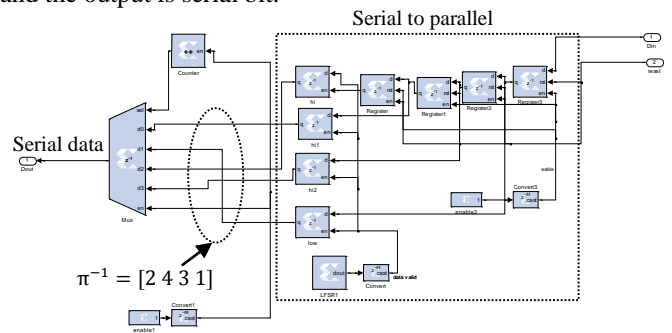


Fig. 20. The serial to parallel converter.

- Serial to parallel

The serial to parallel conversion has been used to convert serial data to parallel two bits streams $d(1:0)$. The slice block has been used to select one bit, the upper slice select $d(0)$ and the lower slice select $d(1)$ as shown in Fig. 21.

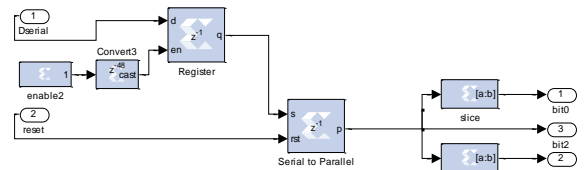


Fig. 21. The serial to parallel converter.

- Viterbi decoder

Viterbi decoder Xilinx IP core version7 has been used to recover information bits as shown in Fig. 22.Viterbi decoder has the same parameter setting of convolution encoder to be consistent with it [15].

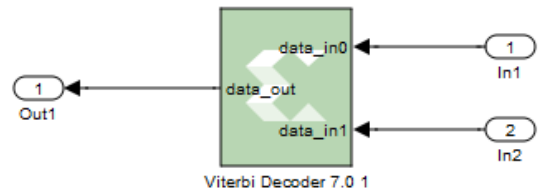


Fig. 22. The Viterbi decoder circuit.

VIII. THE SIMULATION RESULTS

The verification of the implementation has been done via system generator. Fig. 22 shows the time waveform for the error signal between the actual and desired signal during the training mode for $\mu=0.001$ and $\mu=0.006$ respectively. From this Fig. it can be seen that increase the step size (μ) will decrease the time of training mode and converged the signal to the desired value in very fast. In other word increase the step size will decrease the resolution of the recovered signal and hence decreased the performance of the error rate.

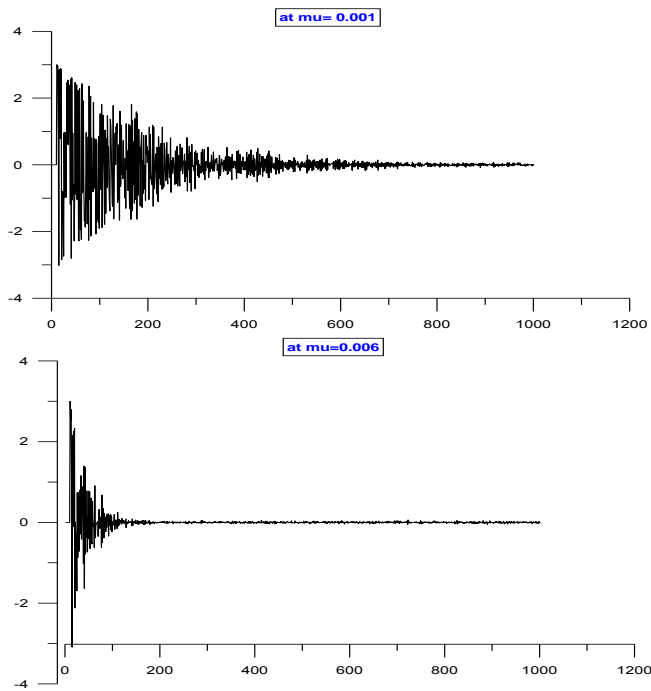


Fig. 23. The time waveform of the error between the actual and desired signals during the training mode for $\mu=0.001$ and $\mu=0.006$ respectively.

Figs. 24 and 25 shows the constellation diagram of 16-QAM for transmitter and receiver side respectively.

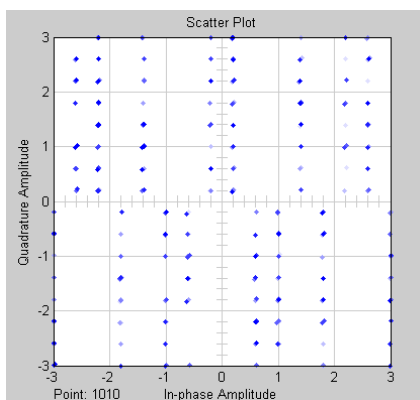


Fig. 24. constellation diagram of 16- QAM at the input of the channel.

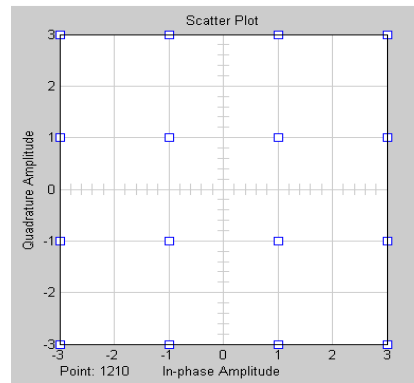


Fig. 25. constellation diagram of 16 QAM at the output of the equalizer.

Fig. 26 and 27 show the simulation results at a transmitter and receiver sides respectively.

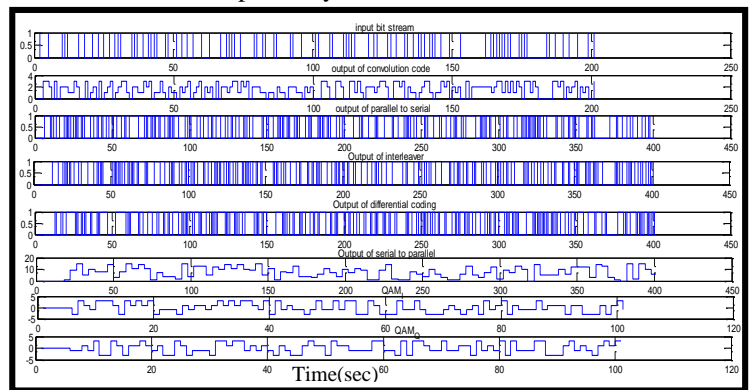


Fig. 26. The time waveforms of transmitter side.

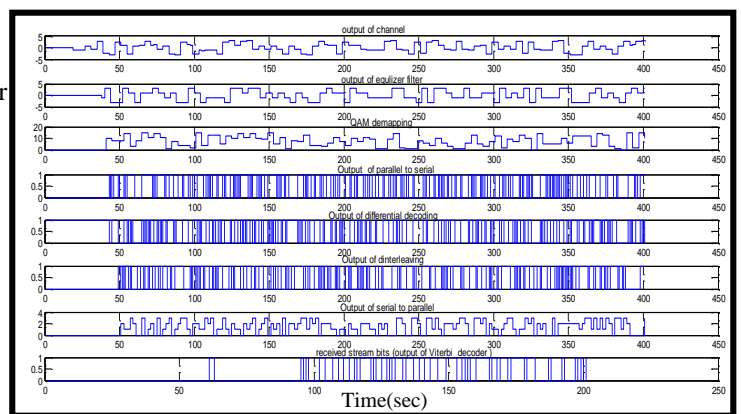


Fig. 27. The time waveforms of receiver side.

Table (I) shows resource utilization and operating frequency.

Table (I) Resource utilization and operating frequency

Device Utilization Summary			
Logic Utilization	Used	Available	Utilization
Number of Slice Flip Flops	2,699	30,720	8%
Number of 4 input LUTs	2,314	30,720	7%
Number of occupied Slices	1,767	15,360	11%
Number of Slices containing only related logic	1,767	1,767	100%
Number of Slices containing unrelated logic	0	1,767	0%
Total Number of 4 input LUTs	2,347	30,720	7%
Number used as logic	1,488		
Number used as a route-thru	33		
Number used as Shift registers	826		
Number of bonded I/Os	4	448	1%
Number of BUFG/BUFGCTRLs	1	32	3%
Number used as BUFPGs	1		
Number of FIFO16/RAMB16s	19	192	9%
Number used as RAMB16s	19		
Number of DSP48s	65	192	33%
Average Fanout of Non-Clock Nets	1.98		

IX. CONCLUSION

In this work a proposed SDR communication system, the system has been designed and implemented using system generator tools to check the effect of using convolution code and efficient LMS linear equalizer. The SDR system generator gives flexibility and optimal in communication system design. The hardware has been implemented on the Xilinx Virtex-4 FPGA using VHDL. A comparison of our proposed work with a conventional LUT-based method and also with a recent work show significant improvement on resource utilization and operating frequency as shown in Table (I). The simulation results show that the LMS filter is good estimation for the channel and the error is zero between the desired and actual signal. Also the results show that the system is synchronized between each component and can be realized in life day as SDR system.

REFERENCES

[1] Chris H. Dick, Henrik M. Pedersen, "Design and Implementation of High-Performance FPGA Signal Processing Datapaths for Software Defined Radios," 2001.

[2] Majid S. Naghmash, Md Hussein Baqir, Mousa Kadhim Wali, "Low Inter Symbol Interference SDR Receiver using LMS Algorithms," *International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 3, no. 2249 – 8958, pp. 410-415, 2014.

[3] S. P. Joshi, Integrating FPGA With Multicore SDR Development Platform To Design Wireless Communication System, Northridge: California State University, 2012 .

[4] N. G. Teena Pahuja, "FPGA Implementation of Adaptive Equalizer for Software Defined Radio," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, no. 12, pp. 334-339, 2013.

[5] C. Manpreet kaur, "Design of an Adaptive Equalizer Using LMS Algorithm," *IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)*, vol. 9, no. 1, pp. 25-29, 2014.

[6] O. P. Sharma, V. Janyani and S. Sancheti , "Recursive Least Squares Adaptive Filter a better ISI Compensator," *International Journal of Electronics, Circuits & Systems*, vol. 3, no. 1, pp. 843-848, 2009.

[7] Raghunandan Swain, Ajit Kumar Panda, "Design of 16-QAM Transmitter and Receiver: Review of Methods of Implementation in FPGA," *International Journal of Engineering and Science*, vol. 1, no. 9, pp. 23-27 , 2012.

[8] Filippo Tosato, Paola Bisaglia, "Simplified Soft-Output Demapper for Binary Interleaved COFDM with Application to HIPERLAN/2," in *IEEE International Conference on Communication (ICC)*, Italy, 2002.

[9] Harpreet Singh Bedi, Gurpreet Singh, Tarundeep Singh, Navdeep Kumar, "Overview of Performance of Coding Techniques in Mobile WiMAX Based System," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 1, no. 2321 – 8169 , pp. 32-35, 2013.

[10] John G. Proakis, and Masoud Salehi , Digital Communication, New York: McGraw-Hil, 2008.

[11] Roberto Garelo, Guido Montorsi, Sergio Benedetto and Giovanni Cancellieri, "Interleaver Properties and Their Applications to the Trellis Complexity Analysis of Turbo Codes," in *IEEE TRANSACTIONS ON COMMUNICATIONS*, 2001.

[12] Robert F.H. Fischer, Lutz H.-J. Lampe, Stefano Calabrò, "Differential Encoding Strategies for Transmission over Fading Channels," *AEU International Journal of Electronics and Communications*, vol. 54, no. 1, pp. 59-67, 2000.

[13] Wang Junfeng, Zhang Bo, "Design of Adaptive Equalizer Based on Variable Step LMS Algorithm," *Proceedings of the Third International Symposium on Computer Science and Computational Technology (ISCST'10)*, vol. 2, no. 1, pp. 256-258, 2010.

Cost Efficient Fast Autonomous Reconfiguration System in Wireless Mesh Networks

N.N.Krishnaveni

Department of Computer science
Research Scholar, Bharathiar University
Coimbatore, India

Dr.K.Chitra

Department of Computer science
Asst.Professor, Govt.Arts College
Melur, Madurai

Abstract— While comparing with the existing networks, Wireless Mesh Network has the advantages of fast implementation, low direct investment and easy maintenance. During their existence time Wireless Mesh Networks (WMNs) experience frequent link failures caused by channel interference, dynamic obstacles, and/or applications bandwidth demands. Wireless mesh networks should be recovered from these link and node failures. These failures cause severe performance degradation in WMNs. In WMNs, the quality of the link can rapidly change because of varying environment condition. The routing algorithm must be able to cope with such changes in link quality and provide alternate route in case the link becomes unusable.

This paper proposes a Cost efficient fast autonomous reconfiguration system (C-FARS) which provides the multiradio Wireless Mesh Networks to recover from link failure automatically to maintain the network performance. C-FARS generates necessary changes in local radio and channel allocations in order to recover from failures. C-FARS recovers from link failures by making cost effective local configuration changes with minimum delay that satisfies the applications QoS demands instead of making global network changes. Our Implementation results shows that C-FARS outperforms existing failure-recover schemes in improving channel-efficiency by more than 95%, also reduces the cost and delay.

Keywords— *Wireless Mesh Networks (WMNs); Cost efficient autonomous reconfiguration system(C-FARS); WMN Architecture, Quality of service (QoS).*

I. INTRODUCTION

A. *Wireless Mesh Networks Architecture*

Wireless mesh networking has emerged as a promising design paradigm for next generation wireless networks. WMNs seem significantly attractive to network operators for providing new applications that cannot be easily supported by other wireless technologies. A WMN is formed by a set of wireless nodes, where each node can communicate and forward data of each other. Wireless Mesh Network consists of two types of nodes: mesh clients (MCs), mesh routers (MRs). Some of the mesh routers act as gateways to the internet using the wired links. These special WMRs are called Internet Gateway (IGWs). During communication the mesh networks divide the long distance into a series of smaller hops to boost the signal using intermediate nodes. Such structural design allows continues flow of data and reconfiguration when paths are blocked or broken. WMN is

fully supported by wireless mesh router network, also called Backbone Wireless Mesh Network (BWMNs). IGWs acts as communication bridges between the internet and BWMN, and provides internet accessibility. These BWMNs provide internet connectivity to MCs.

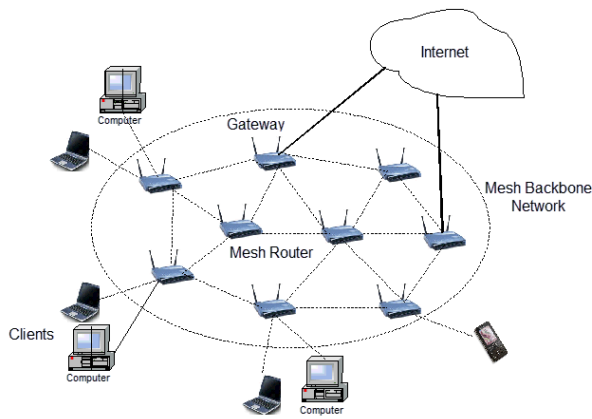


Fig. 1. WMNs architecture

B. Why Is Self-Reconfigurability Necessary?

Wireless Mesh Network has the advantages of fast implementation, easy maintenance and low direct investment while comparing with the existing networks. During their life time Wireless Mesh Networks (WMNs) experience frequent link failures caused by channel interference, dynamic obstacles, and/or applications bandwidth demands. Wireless mesh networks should be recovered from these link and node failures. These failures cause severe performance degradation in WMNs.

The topology and the connectivity of the network can vary frequently because of route failure and energy depletions; an efficient self-configuration, topology control and power managements are required. In WMNs, the quality of the link can rapidly change because of varying environment condition. The routing algorithm must be able to cope with such changes in link quality and provide alternate route in case the link becomes unusable. This approach gives the research on the fault-tolerance for Wireless mesh networks

C. Techniques For Link Recovery In WMN

Though WMNs are used widely, they often face frequent link failures. So link recovery plays an important role in WMN. There are various techniques used for link recovery in WMN.

1) *Initial Resource Allocation Method*: In this method, using some theoretical guidelines for allocation of resources, the initial planning is done to recover from link failure. Even the planning is done before, this method has drawback of "Global reconfiguration changes". By using this method for link recovery where a small change is required, this method performs reconfiguration of the entire network [3].

2) *Greedy Channel Assignment Method*: In this method the drawback of "Initial Resource Allocation Method" is removed. In this method, the setting of faulty link is alone changed instead of the entire network. Even it do local changes it too has drawback of "ripple effect". In this effect

whenever a local change is done then it causes triggering of change to some kind of network settings [4].

3) *Fault Tolerant Routing Protocol* : When there is a link failure during the communication in WMN, then it is must to recover the link from failure and also make sure that the communication must not loss. To overcome these link failures and avoid the data loss we can route the packet through different link. Protocols help us in selecting the alternate paths. Protocols such as rerouting or multipath routing can be used. Anyway these protocols are not that much efficient and take massive amount of time to reroute, which leads to delay delivery of packets.

4) *Autonomus Reconfiguration System (ARS)*: Kim and Shin [5] proposed a new recovery technique for wireless mesh networks named ARS. In this method the node detects the link failure and generates a set of reconfiguration plans by considering the rage of channels. Among the set of plans, a feasible plan is selected by considering that the plan must maximizes the throughput of the network and also satisfy some QoS constraint of the network. The drawback of ARS is that it is not cost aware reconfiguration technique.

5) *Enhanced Reconfiguration System (ERS)*: Ramakrishnan R and Dr. N.Sankar introduces ERS [6] is to provide cost aware reconfiguration system in wireless mesh networks. In ERS a set of reconfiguration plans are generated and among them a best plan is selected which provides the required service in minimal cost. According to the selected plan the reconfiguration is done in the network. Thus it maximizes the network performance. The drawback of ERS is, sometimes it may fail to satisfy the Qos constraint which leads to frequent link failure.

6) *Quick Autonomous Reconfiguration System (QARS)*: QARS [7] is proposed by A.Melveena & D.Ramya Dorai to recover the link failure in a short duration to avoid the delay transmission. In this method when a link failure is detected, immediately a group is formed around the faulty area. Among the group, a leader node is selected using enhanced Bully algorithm. Using the plans generated by other nods in the group, the leader node select the plan which requires minimum change and meets the Qos demand. Thus it improves the network performance.

II. C-FARS ARCHITECTURE AND ALGORITHM

A. C-FARS Architecture

In the proposed system a network is assumed to consist of mesh nodes, IEEE 802.11 based wireless links and control gateways as in fig. 2.

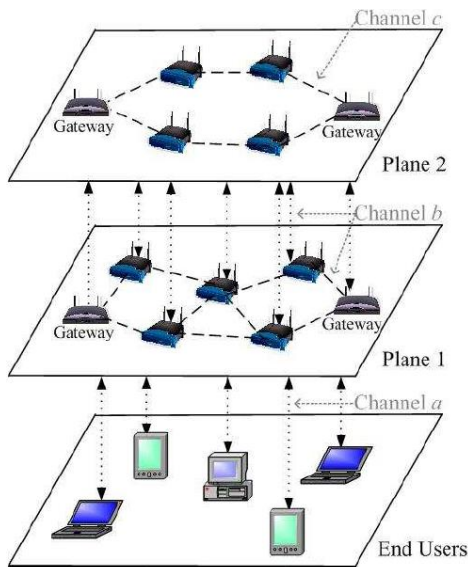


Fig. 2. C-FARS architecture

C-FARS running in every mesh node supports self-reconfigurability via the following distinct features:

- *Localized Reconfiguration:* Based on the multiple channels and radio associations available, C-FARS generates reconfiguration plans that allow for only local network changes where the link failures occurred.
- *Fast Reconfiguration:* C-FARS generates the reconfiguration plans within minimum duration and reduces the delay in reconfiguration. Such fast reconfiguration helps in increasing network performance.
- *QoS-Aware Planning:* C-FARS effectively identifies QoS-satisfiable reconfiguration plans by : 1) Estimating the QoS-Satisfiability of generated reconfiguration plans, and 2) Deriving their expected benefits in channel utilization.
- *Cross-Layer Interactions:* C-FARS actively interacts across the network and link layers for planning. This interaction enables C-FARS to include rerouting for reconfiguration planning in addition to link-layer reconfiguration. C-FARS also maintains connectivity during the recovery period with the help of a routing protocol.
- *Cost Effective Multipath Selection:* The routing algorithm must be able to cope with changes in link quality and rapidly provide an alternative route in case the link becomes unusable. Our modified DSR algorithm in C-FARS selects the least cost alternative path based on various routing metrics among the multiple paths available in case of path failure.

B. C-FARS Algorithm

In this paper we proposed an algorithm for C-FARS which operates on each and every node of the Wireless Mesh Networks (WMNs). This algorithm executes in a particular time interval to monitor the failure and if found it recovers from such failure in an efficient manner.

Using this algorithm it recovers from the link failure and if there is a frequent link failure detected on the same link which means the particular link unusable, then it is consider as a path failure and a new cost effective path is selected using modified DSR.

-
- (1) Monitoring period
 - 1: **for every link do**
 - 2: measure link-quality using passive monitoring;
 - 3: **end for**
 - 4: send monitoring results to a gateway ;
 - (2) Failure detection and group formation period
 - 5: **if** link violates link requirements **then**
 - 6: request a group formation on channel of link;
 - 7: **end if**
 - 8: participate in a leader election if a request is received;
 - (3) Planning period
 - 9: **if** node is elected as a leader **then**
 - 10: send a planning request message to a gateway;
 - 11: **else if** node is a gateway **then**
 - 12: synchronize requests from reconfiguration groups
 - 13: generate a reconfiguration plan;
 - 14: send a reconfiguration plan to a leader;
 - 15: **end if**
 - (4) Reconfiguration period
 - 16: **if** includes changes of node **then**
 - 17: apply the changes to links;
 - 18: **end if**
 - 19: relay to neighboring members, if any

Algorithm. 1.

C-FARS algorithm 1 involves 4 phases

1) *Network Monitoring:* In this 1st phase each mesh node monitors the quality of its outgoing wireless links at every t_m sec (eg.10 sec) and the status of the results are reported to the gateway.

2) *Failure Detection Phase And Group Formation:* System detects the failure by comparing the current link state information with that in the existing database. Whenever a failure is detected all the mesh nodes that use a faulty channel are grouped together.

3) *Planning Period:* After forming a group one of the group members will be elected as a leader by using the well-known bully algorithm and this leader node will send a plan request message to the gateway. Now the gateway generates the new reconfiguration plan which is broken down into 3 steps as in fig. 3.

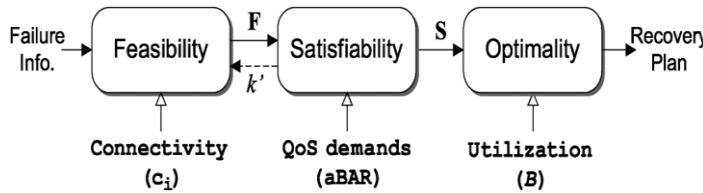


Fig. 3. Steps in planning period

a) *Feasible Plan Generation:* Generating feasible plans is very important to keep in mind all legitimate changes in links configurations and their combinations around the faulty area. C-FARS identifies feasible changes that do only local changes but maintain existing network connectivity as much as possible. While selecting feasible plan C-FARS will 1) avoid a faulty channel. 2) Maintain network connectivity and utilization. 3) Control the scope of reconfiguration changes.

b) *QoS-Satisfiability Evaluation:* Among a set of feasible plans, C-FARS filters the plan that satisfying QoS by checking if the QoS constraints are met under each plan. C-FARS will select the plan by 1) Per-link bandwidth estimation 2) Examining per-link bandwidth satisfiability 3) Avoiding cascaded link failures

c) *Choosing the Best Plan:* C-FARS now has a set of reconfiguration plans that are QoS-satisfiable and needs to choose a plan within the set for a local network to have evenly distributed link capacity. C-FARS selects the plan by 1) Quantifying the fairness of a plan 2) Path with the highest minimum access efficiency value 3) Breaking a tie among multiple plans

4) *Implementing Reconfiguration Plan:* Here firstly the gateway sends the selected reconfiguration plan to the leader node then leader node distributes it to all other nodes in the group and then each node executes the corresponding configuration changes

III. PATH SELECTION IN C-FARS

C-FARS is proposed to overcome the failure by making the reconfiguration with minimum delay and less expensive. Once a node detects the failure, it self-reconfigures by using feasible plans. If there is no feasible plan for a particular link failure or it detects frequent failure on a particular link, an alternate route must be discovered. The new route selected must be cost effective, feasible, optimal, and satisfies QoS demands.

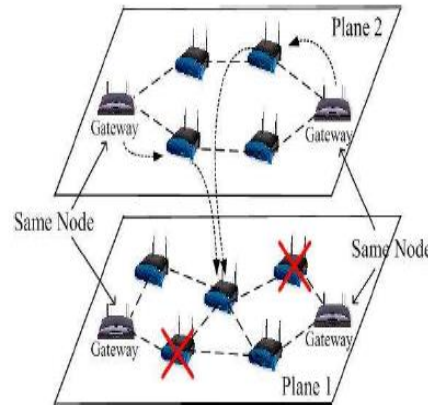


Fig. 4. C-FARS Failure and reconfiguration

A. Route Discovery Using Modified DSR

The rationale for modifying the DSR protocol is to make it better suited to the WMN environment. The WLAN medium is a shared medium where nodes must contend for accessing the medium using DCF MAC mechanism. Since the DCF is a “listen before talk mechanism”, a high level of contention for access to the medium will result in a low availability of bandwidth at a node. This in turn limits the maximum throughput that can be achieved. Unfortunately, the DSR protocol fails to explicitly consider the availability of bandwidth locally at a node which is an important omission in WMNs based upon the IEEE 802.11 standard.

In this case, the access efficiency measured locally at a node is used as a measure of the local availability of bandwidth at that node. By incorporating bandwidth availability information into the DSR protocol the cost of Route Discoveries can be reduced and the overall performance of the network can be significantly improved. The performance of modified DSR is investigated through a series of simulations performed on the NS2 modeler package.

In this work a new metric to support the DSR node cache mechanism has been used. The modification to the DSR protocol is intended to incorporate knowledge of the path capacity into the route discovery mechanism. Specifically, the DSR protocol was modified by replacing hop count field in the cache route table with an access efficiency field. The

optimal route is determined by selecting the path with the highest minimum access efficiency value.

Fig. 5. Illustrates the operation of the path selection mechanism for the DSR protocol and our modified DSR protocol. From this figure, the original DSR protocol selects path B as the hop count of this path is smaller than the hop count of the other paths (path A and path C). While our modified DSR protocol chooses path A over paths B and C as it selects the path with the highest minimum link capacity.

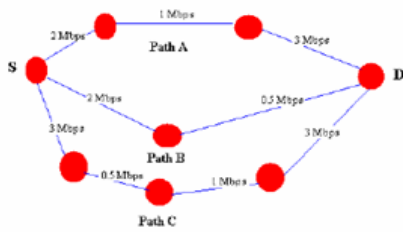


Fig. 5. Path selection in C-FARS

B. Path Selection Metric

In this section we present a new interference-aware multipath selection metric. Our selection metric and algorithm aims to minimize interference (between paths and to neighbor nodes), without assuming that interference is global, and considers link quality and the delaying effect of long paths, when selecting paths.

1) *Path Interference Cost*: The path interference cost reflects the degree of interference between links operating on a common channel along the selected paths. To help define the path interference cost, we first define the interference cost for a link (i,j) on channel c in a network N as:

$$LI_{ij}(c, N) = ETT_{ij}(c) * S_{ij}(c, N), \quad (3)$$

Where $S_{ij}(c, N)$ denotes the number of nodes in network N that are affected by interference from link (i,j) on channel c . In this work we assume interference to be binary – two nodes either do or do not interfere. This assumption allows us to simplify the computation of metrics. We plan to investigate more accurate classifications of interference in the future. The path interference cost for a set of paths P is simply the aggregate of all link interference costs along the paths:

$$PIC_P = \sum_{ij \in P} LI_{ij}(c, P) \quad (4)$$

2) *Neighbour Interference Cost*: Similar to the path interference cost, the neighbor interference cost represents the

channel time cost to nodes close to the paths, and it is defined as:

$$NIC_P = \sum_{ij \in P} LI_{ij}(c, N - P), \quad (5)$$

3) *Weighted Interference Multipath metric*: The Weighted Interference Multipath (WIM) metric is a weighted average of path interference and the neighbor interference costs.

$$WIM_P = \beta * NIC_P + (1 - \beta) * PIC_P, \quad (6)$$

where the parameter β satisfies $0 \leq \beta \leq 1$.

The WIM metric can be interpreted as a balance between local and global considerations. The path interference cost component reflects the total channel time along the paths that is consumed when the channels are concurrently used. The smaller it is, the better the paths will be at providing low end-to-end delay, assuming interference to neighbor nodes is negligible. On the other hand, the neighbor interference component favors paths that have less interference to nodes that are not on the paths. This could be beneficial when the network load is high such that interference from neighboring nodes starts to affect, or even dictate, the traffic on the paths.

C. C-FARS Fault Tolerance

Physical C-FARS can play very significant role for failure recovery in a WMN. As any node in plane 1 is reachable from plane 2 in minimum hops, if any node on plane 1 fails, all the data destined to that node can be redirected over plane 2. Moreover, because of the ring architecture, if any node on C-FARS suffers a failure, data can still be routed over other part of the C-FARS. As all the gateways of plane 1 is included in plane 2, if a gateway fails, then traffic can be redirected to the other gateway, and C-FARS can carry the traffic from the other gateway, and deliver the packets. Failure recovery schemes described can be integrated with C-FARS for protection. Fig. 4. shows how C-FARS can be used for fault tolerance in a WMN.

D. Energy Efficient Model

The topology control problem is a well researched topic for energy saving in wireless ad hoc networks. However, little attention has been given to similar problems in the case of wireless mesh networks (WMNs) even though WMNs have very unique characteristics that are different from other wireless multihop networks e.g., MANETs. This is because many WMN surveys make the impractical assumption that since mesh routers are static, energy is not a problem.

Consequently, with specific interests to WMN applications in rural areas, where power sources are limited, this work addresses the topology control problem for energy efficiency in a hybrid WMN of heterogeneous wireless devices with

varying transmission ranges. A localized distributed algorithm is presented which computes an optimal per-node transmission power such that: (1) a node's average out degree is reduced considerably to cover only the nearest neighbors, (2) network connectivity is maintained and (3) the network lifetime is extended. The performance of the algorithm is evaluated via several mathematical analyses.

Additionally, simulations are done in the NS-2 simulation environment to show correctness and effectiveness of the algorithm. A cross-layer modification to the DSR protocol that increases the global throughput in wireless mesh networks. In our modified DSR protocol we have introduced the Access Efficiency metric as an alternative to the hop-count metric in order to improve the route selection mechanism.

The selected path in the route selection mechanism is identified by choosing the path with the highest minimum Access Efficiency value. We have employed the NS2 modeler as a simulator to examine two different patterns of traffic for a series of randomly generated network topologies. Each topology was simulated twice. One simulation used the original DSR algorithm while the other utilized the modified DSR algorithm. The average throughput was recorded for each run and the percentage improvement for the particular topology was calculated. Our results demonstrate that a significant increase in the global throughput of the networks can be achieved by implementing our modified DSR protocol.

IV. PERFORMANCE EVALUATION

Our following experimental evaluations on an ns2-based simulation will demonstrate the effectiveness of C-FARS in recovering from local link-failures and in satisfying applications' diverse Qos demands

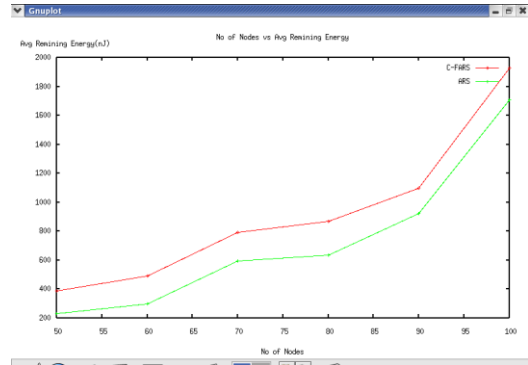


Fig. 8. Average remaining energy

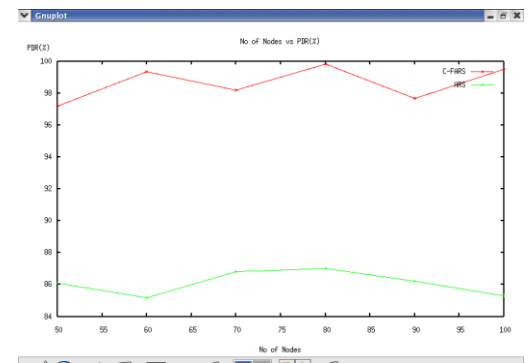


Fig. 9. Packet delivery ratio

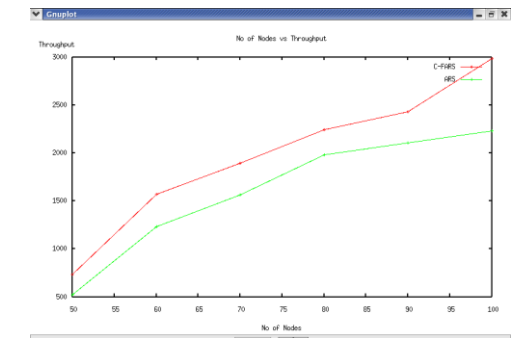


Fig. 10. Throughput

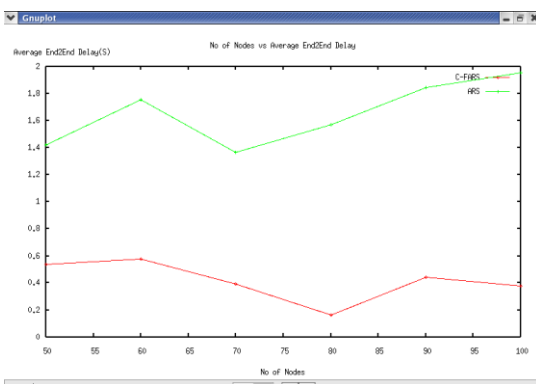


Fig. 6. Average end-to-end delay

Fig. 7.

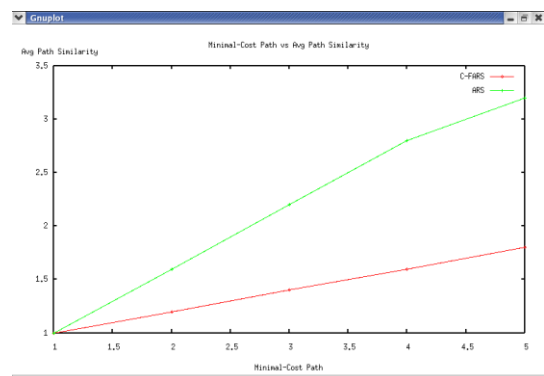


Fig. 11. Minimal cost path (path metric)

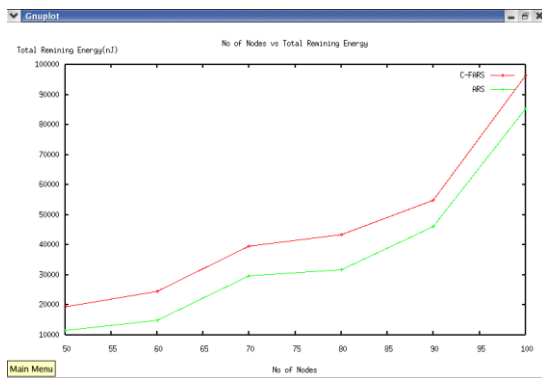


Fig. 12. Total remaining energy

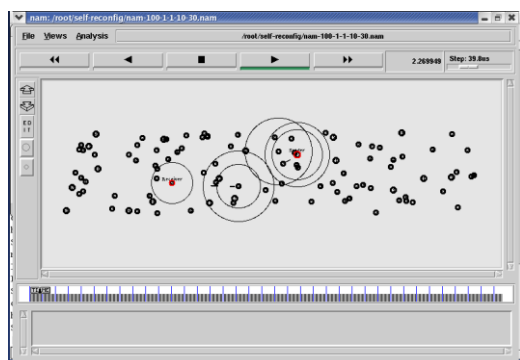


Fig. 13. Data transfer

V. CONCLUSION

This paper proposes a Cost efficient fast autonomous reconfiguration system (C-FARS) which provides the multiradio Wireless Mesh Networks to recover from link failure automatically to maintain the network performance. C-FARS generates necessary changes in local radio and channel allocations in order to recover from failures. C-FARS recovers from link failures by making cost effective local configuration changes with minimum delay that satisfies the applications QoS demands instead of making global network changes. Our Implementation results shows that C-FARS outperforms existing failure-recover schemes in improving channel-efficiency by more than 95%, reduces the cost and delay.

. This scheme improves the reliability of dynamic wireless sensor networks in the point-point routing scenario by using multipath routing. This method is suitable to disseminating a large amount of bulk data to the destination with a high reliability and low delay.

Our cross-layer approach achieves the proposed reliability improvement in a dynamic wireless sensor network. Our lessons learned in this research show that in the routing

protocols for WSN, optimization is more effective when taking into account the overall system and with the use of all available knowledge, instead of a strict layered approach. This approach has a much larger network lifetime, compared with traditional protocols for WMNs.

REFERENCES

- [1] A.Valarmozhi, M.Subala, V.Muthu, "Survey of Wireless Mesh Network", International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 6, December 2012
- [2] Mojtaba Seyedzadegan, Mohamed Othman, Borhanuddin Mohd Ali and Shamla Subramanian, "Wireless Mesh Networks: WMN Overview, WMN Architecture", International Conference on Communication Engineering and Networks, 2011, IPCSIT vol.19.
- [3] Alicherry M, Bhatia R, and Li L, "Joint channel assignment and routing for throughput optimization in multi-radio wireless mesh networks," in Proc. ACM MobiCom, Cologne, Germany, pp. 58–72, Aug. 2005.
- [4] Raniwala A and Chiu T, "Architecture and algorithms for an IEEE 802.11-based multi-channel wireless mesh network," in Proc. IEEE INFOCOM, Miami, FL, Vol. 3, pp. 2223–2234, Mar. 2005.
- [5] Kyu-Han Kim and Kang G. Shin, IEEE/ACM TRANSACTIONS ON NETWORKING, "Self-Reconfigurable Wireless Mesh Networks", APRIL 2011.
- [6] Ramakrishnan R, Dr. N. Sankar Ram, Dr. Omar A. Alheyasat, "A Cost Aware Reconfiguration Technique for Recovery in Wireless Mesh Networks", IEEE/ICRTIT, 2012.
- [7] A.Melveena, D.Ramya Dorai, "QARS for Self Reconfiguration Mechanism in Wireless Mesh Networks", International Journal of Advanced Research in computer Engineering & Technology (IJARCET) Volume 2, Issue 2, February 2013
- [8] Kalyani Pendke 1 and S.U.Nimbhorkar, "STUDY OF VARIOUS SCHEMES FOR LINK RECOVERY IN WIRELESS MESH NETWORK", International Journal on AdHoc Networking Systems (IJANS) Vol. 2, No. 4, October 2012.
- [9] B.Lakshmi Sowmya1, M.Srinivas 2, B.Sravani3, "A NOVEL METHOD OF DYNAMIC SELF RECONFIGURABLE WIRELESS MESH NETWORK", International Journal of Engineering Research and Development e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com Volume 4, Issue 1 (October 2012).
- [10] K.H. Kim and K. G. Shin, "On accurate and asymmetry-aware measurement of link quality in wireless mesh networks," IEEE/ACM Trans.Netw., vol. 17, no. 4, pp. 1172–1185, Aug. 2009.
- [11] P. Kyasanur and N. Vaidya, "Capacity of multi-channel wireless networks: Impact of number of channels and interfaces," in Proc. ACM MobiCom, Cologne, Germany, Aug. 2005, pp. 43–57.

PARAMETER ANALYSIS FOR CLUSTERING IN MANET IN DISASTER SCENARIOS

V.Preetha

Research Scholar,
Bharathiar University,Coimbatore,
Tamilnadu,India
preetha_mca2005@yahoo.com

Dr.K.Chitra

Assistant Professor,Dept.of Computer science
Govt.Arts College,Melur
Madurai,Tamilnadu,India

Abstract— Mobile Adhoc Network has become an essential one in every aspects of our life due to the recent growth of technological developments. Effective communication in right scenarios is an important factor to be considered. Here in this paper, Cluster based routing for a specific disaster scenario and the parametric analysis for the effective clustering in Mobile adhoc network (MANET) is analyzed and proposed using fuzzy logic with mat lab simulator.

Keywords— MANET,Hierarchical routing,Cluster based routing schemes for rescue scenarios,fuzzy logic

I. MOBILE ADHOC NETWORK

A mobile ad hoc network consists of a group of wireless-enabled devices. Initially, it was named as packet radio network and was initiated by the Department of Defense (DoD) of the United States of America. Mobile adhoc network itself is an autonomous system consisting of collection of mobile hosts connected by wireless links. It has no supporting fixed infrastructure or central administration and the nodes will communicate within the transmission range of each other. If two hosts are unable to communicate, then they will try to communicate if other hosts lying in between are willing to forward packets for them. Thus every node will participate in multi-hop routing to reach all the nodes in the network. Continuous improvement in smart devices in wireless world has gained the interest of the users in capitalizing these smart devices. MANET is widely used in rescue operations, military scenarios and in scenarios where it is impossible to establish a wired backbone. While considering the emergency situation particularly in mass disasters, a quick and coordinated response must be given to improve the efficiency of rescue teams and to save as many lives as possible. Furthermore, the emergency situation may be ongoing for some time or even days; hence systems may have to stay usable for extended periods. This is the case of IMPROVISA (Improvisa) that proposes to solve this difficulty by distributing antennas in the disaster area [1]. Some other applications of MANET include the adhoc network created in smart classrooms by the students and Professors with the laptops and by a group of people at a

meeting using their PDAs or laptops to exchange information among themselves. In future, MANET will be widely used in ubiquitous computing, providing connectivity to everyone, anywhere and from any device. As a result, robust routing mechanisms and improving the Quality of service becomes inevitable. Mobility of nodes in MANET and limited battery power, dynamic topology changes, link bandwidth are major challenges in MANET when concerned with routing, scalability and management functions.

II. HIERARCHICAL ROUTING IN EMERGENCY SCENARIOS

A. Routing in MANET

Mobile adhoc network undergoes multihop network topology that may change frequently due to mobility, congestion in traffic, power constrained problems and computational overhead. Many routing protocols are set up for efficient routing. Proactive routing, reactive routing, Hybrid routing protocols and the route constructed are kept alive as long as possible. Location aided routing protocol is based on the nodes location information. The communication overhead of proactive protocols is $o(n^2)$ where n represents the total number of mobile terminals [2].Reactive routing protocols faces scalability and mobility challenges. Hybrid protocols exhibit both reactive and proactive features. The drawbacks of proactive and reactive protocols such as overheads and delay are minimized in Hybrid routing. A hierarchical routing depends on the hierarchic level in which a node resides. For

the better performance of MANETs a hierarchical architecture is essential.

B. Emergency Rescue Scenarios

MANET is widely used in Emergency rescue scenarios, Disaster recovery, supporting doctors and nurses in hospitals, Environmental disasters etc. Hence in such cases continuous evaluation of the routes between the routes is very important to rescue the people affected. In case of emergency and rescue scenario, there may be obstacles which affect the normal mobility pathways. Many network providers are providing instant networks in disaster areas with mobile base stations, antennas and power generators. These instant networks will be life saving because the deployment of placing a fixed network will take much duration and the cost will be high. When Typhoon Bopha hit the Philippines an instant network was created and many people were rescued. While analyzing the emergency scenarios, zone based link connectivity and mobility are the important factors to be considered. In order to overcome the disadvantages of reactive and proactive hierarchical structure has been developed. In general, Cluster based routing protocol [CBRP] falls under the reactive category but due to its level-oriented administration and governance by cluster head it has the hierarchical component. CBRP has many advantages such as energy consumption and network performance. Thus a typical hierarchical structure can be implemented by partitioning the network into clusters depending on the geographic region, transmission range, and communication reliability irrespective of the sparse and dense regions.

III. CLUSTERING SCHEME

A. Cluster based routing for Rescue Scenario

Clustering refers to the partitioning of the network in to different sizes. Transmission range and Zone based partitioning will aid in the easier routing. Liliana Enciso Quispe et.al analyzed the hierarchical routing in their work based on the Behavior of Ad Hoc routing protocols, analyzed for emergency and rescue scenarios, on a real urban area[3] and suggests that the use of CBRP protocol in a disaster area, more efficiently adjusts the evacuation of persons and their care and appropriate location. The following figure [Fig.1] is the scene for the Emergency rescue scenario of the City Loja with an area of 1000m x500m. The diagram represents the Node or a person with a mobile device as a cyan circle, goal point as the red star, and obstacles with the red circles with a black stripe in the middle. The group leader is represented as the red circle.

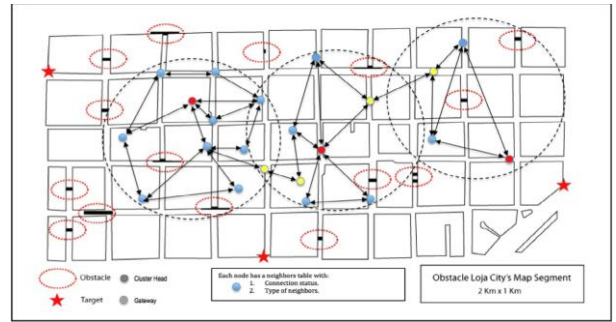


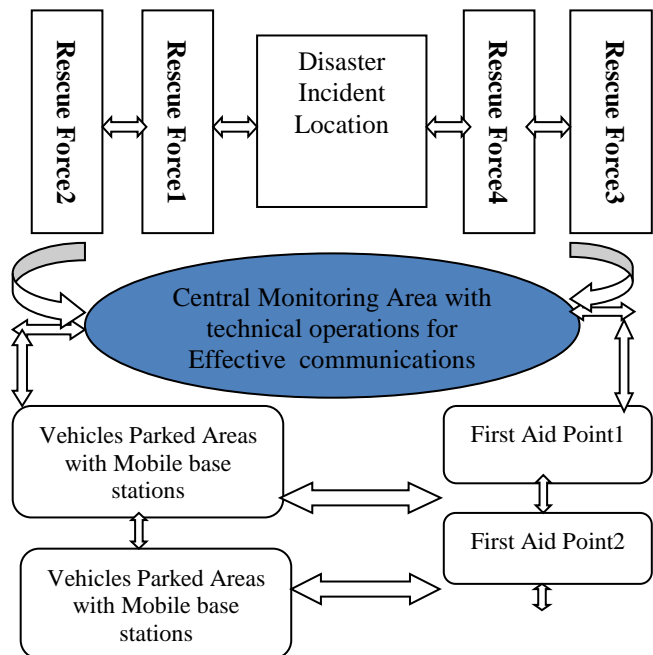
Fig. 1. This is an example of Cluster Based Adhoc network.

CBRP routing will be very efficient in the case of rescue operations for optimum routes, large scalability, traffic reduction etc.

IV. NETWORK MODEL AND PROBLEM SPECIFICATION

A. Disaster area model

This paper focuses on the following disaster area model with network connectivity. This temporarily created MANET involves mainly communication devices for effective communication. This may include even vehicular adhoc network, Air network, fly network with mobile base stations. Clustering in this type of MANET will be very suitable for effective communication. This may save the energy of the resources and for fastest telecasting and monitoring of the current situation. The Government Hospitals, Police station and Telecasting media centre may be at a particular distance. Connectivity index can be used to find the shortest path.



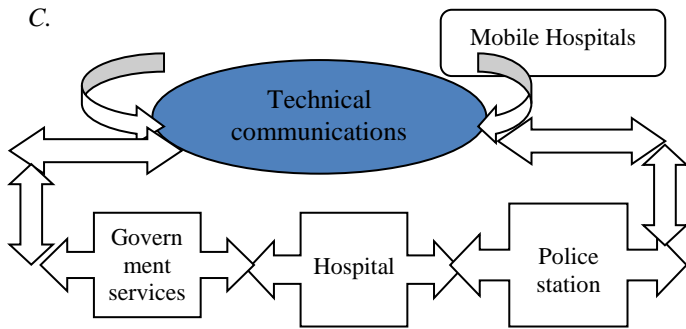


Fig.2. Disaster Area Model

We are analyzing some of the following necessary parameters for the effective chance of clustering in MANET.

V. PARAMETER ANALYSIS USING FUZZY LOGIC

A. Connectivity Index

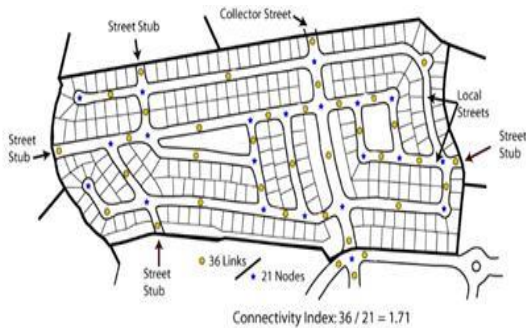


Fig.3. Connectivity Index of the City of Fayetteville, North Carolina

The above Connectivity index represents the area in the the City of Fayetteville, North Carolina [4]. Connectivity Index is also Known as Randić index. It was proposed by Randić in 1975.[5] $\chi(G)$ is defined as the connectivity index and it was calculated based on the degree of vertex. In general, Connectivity index is also an important entity when analyzing the MANET implementation in the above disaster scenarios. Nodes (stars) exist at street intersections. Links (circles) are represented as stretches of road that connect nodes. In the above diagram, there are 36 links (circles) and 21 nodes (stars); therefore the connectivity index is 1.71 ($36/21 = 1.71$). The greater connectivity index factor shows the links between the source and destination for easier reachability and routing. By assuming that if the disaster area is divided into zones as below:

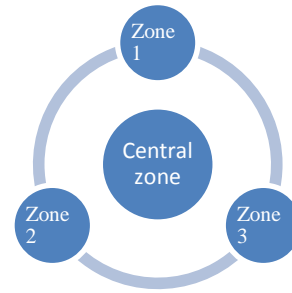


Fig.4. Disaster Area Divided in to Zones.

If zone1 and Zone2 recovery process was finished, the links can be added or deleted and connectivity can be redirected. In some Connectivity analysis problems, the connectivity is defined by the probability that a node is reachable at any other node in the network. For a single component graph, any node is reachable at other nodes, thus, the connectivity is equal to 1. So connectivity can be defined as

$$Connectivity = \frac{Largest\ Connected\ Component}{N} \quad (1)$$

Where N is the total number of nodes in the network.

B. Mobility

For this scenario, the assumption is that the mobility of the nodes will not be too fast. It depends on the incident area zone. Many mobility models have been described by authors. Some of them are real life mobility models, models based on topology restrictions and statistical models with random mobility. Different mobility metrics with relative speed velocity are calculated based on the position of the nodes. But in a Disaster area scenario GPS system will be helpful for identifying the position of the nodes even though the hardware cost is high and the energy consumption is high.

C. Transmission range

The transmission range may vary from short range to long range. The disaster scenarios will have the transmission range depending on the incident location and the density of the nodes. A larger transmission range will have greater connectivity. The total number of nodes within the transmission range is defined as :

$$\sum_{u \in v, u \neq v} \{D_{uv} < T_x, Transmission\ range\} \quad (2)$$

D. Density of the nodes

If the density of the nodes is high there may be higher attenuation. Generally, the density of the nodes in the disaster area may depend on the population of the area.

E. Battery energy

Each mobile node has a defined amount of energy. The energy of the node is the important factor in communication devices. If this amount is low, the node can neither send nor receive any data. The energy consumption of a node during a network will have the following states :

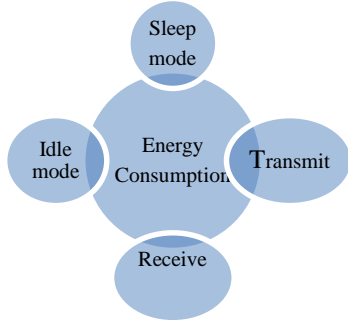


Fig.5.Energy States

Here in this disaster scenario, we are not considering the energy of a particular node to calculate the membership function. But the leader node has to be chosen for clustering based on highest energy.

F.Fuzzy Logic

The fuzzy logic was used to represent uncertainties. Fuzzy logic control system is rule-based system which contains a set of fuzzy rules. It will be very useful for control decision mechanism. Mamdani fuzzy-rule based systems has the linguistic description which has both the antecedent parts and the consequent parts. Rule base is an IF-THEN rule group with fuzzy sets that represents the desired behavior of a fuzzy system.

$$IF x_1 \text{ is } A_{i1} \text{ and } \dots \text{ and } x_n \text{ is } A_{in} \text{ THEN } y \text{ is } C_i, i = 1, 2, \dots, L \quad (3)$$

Where L represents the number of fuzzy rules, X is the input variable, y is the output variable. A_{ij} are the fuzzy sets of the input linguistic variable x_j and c_i represents the set of the output linguistic variable y. A_{ij} and c_i are characterized by both membership functions. For the above discussed parameters, the fuzzy membership function is calculated and the effect of clustering in the above discussed disaster scenario MANET is discussed by rule-based system.

TABLE I

SELECTED PARAMETERS FOR CLUSTERING WITH RANGE

Parameters Chosen for Clustering	Range
Connectivity Index of a particular Zone	High
	Medium
	Low
Transmission Range	Long
	Short

	Medium
Mobility	High
	Medium
	Less
Density	Large
	Medium
	Small

The membership functions of the above parameters are derived using fuzzy logic in Mat lab as:

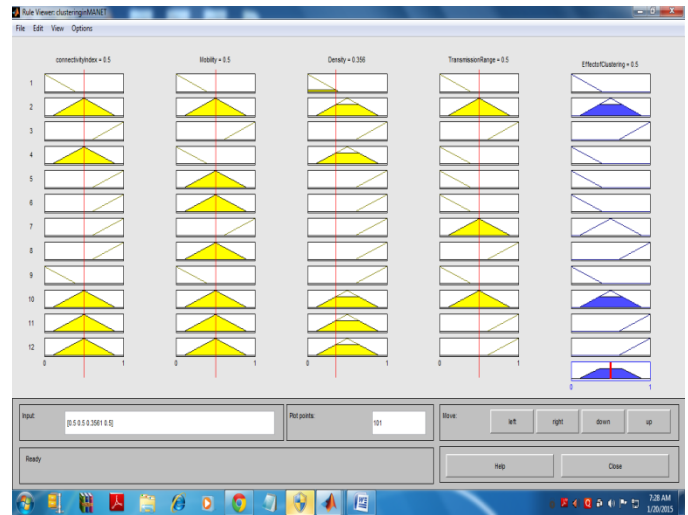


Fig.6.Membership functions for the selected parameters

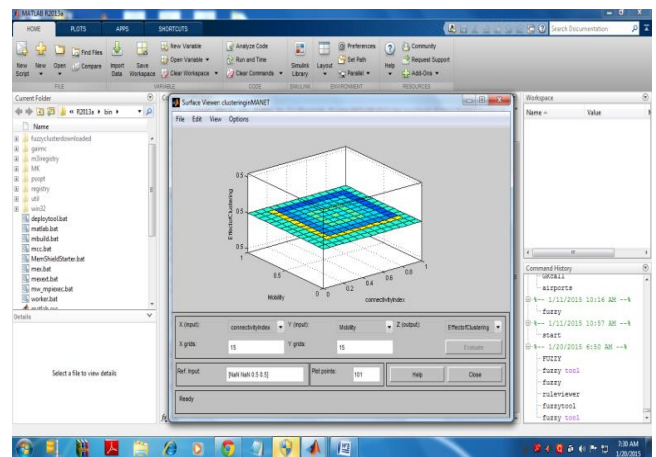


Fig.7.Parameter Analysis Graph for effective clustering

Thus the required important parameters for the clustering in MANET in case of disaster scenarios is discussed using fuzzy logic. The rules are generated for the chance of effective clustering in MANET using Rule-based Mamdani. This shows that if the parameters connectivity index and the transmission

range is high and if the mobility and density are medium then the chance of the effective clustering is best. The graph shows the impact of the parameters on the effective clustering for MANET.

VI. CONCLUSION

This paper focuses on the required important parameters for the effective clustering chance especially in focus of the disaster scenarios. It proves that the chance of effective clustering is possible in disaster scenarios and the parameters are analyzed by focusing on efficient clustering. In future, the effective head will be chosen depending on the proposed parameter values.

REFERENCES

- [1] Abraham Martin-Campillo, Abraham, et al. Jon Crowcroft, Eiko Yoneki, Ramon Marti, "Evaluating opportunistic networks in disaster scenarios," *Journal of Network and Computer Applications*, 2013, pp. 870–880.
- [2] Jane Y. Yu and Peter H.J. Chong, "A Survey of Clustering Schemes for Mobile Adhoc networks", *IEEE Communications Surveys & Tutorials*, vol. 7, 2005, pp. 32–47
- [3] <http://www.cityoffayetteville.org/vic/default.htm?url=Documents%2F305f4f2connectivityi.htm>
- [4] M.A. Rajan, M. Girish Chandra, Lokanatha C. Reddy and Prakash Hiremath, "A Study of Connectivity Index of Graph Relevant to Ad Hoc Networks", *IJCSNS, International Journal of Innovative Research in Computer and Communication Engineering*, vol. 2, 2014, pp. 5152–5157.
- [5] Liliana Enciso Quispe, Luis Mengual Galan, "Behavior of Ad Hoc Routing protocols analysed for emergency and rescue scenarios, on a real urban area", *Expert systems with applications*, 41.5 (2014): 2565–2573.
- [6] Sharmila John Francis and Elijah Blessing Rajsingh, "Performance Analysis of Clustering protocols in Mobile Adhoc Networks", in *Journal of Computer Science* 4 (3), 2008, pp. 192–204
- [7] Mainak Chatterjee, Sajal K. Das and Damla Turgut, "WCA: Weighted Clustering Algorithm for Mobile Adhoc networks", in *Cluster Computing* 5, 2002, pp. 193–204.
- [8] Charalampos Konstantopoulos, Damianos Gavalas, Grammati Pantziou, "clustering in Mobile Adhoc network through neighbourhood stability based mobility prediction", in *Computer Networks*, 2008, pp. 1797–1824.

Extending OpenFlow in Virtual Networks

Lorena Isabel Barona López, Ángel Leonardo Valdivieso Caraguay, Luis Javier García Villalba

Group of Analysis, Security and Systems (GASS)
Department of Software Engineering and Artificial Intelligence (DISIA)
Faculty of Information Technology and Computer Science, Office 431
Universidad Complutense de Madrid (UCM)
Calle Profesor José García Santesmases, 9
Ciudad Universitaria, 28040 Madrid, Spain
Email: {lorebaro, angevald}@ucm.es, javiergv@fdi.ucm.es

Abstract— Software Defined Networking (SDN) is a novel technology that has become a prominent topic in the last years. In any research is essential to have emulators and simulators in order to test new applications or protocols. In this context, we present the integration of OpenFlow protocol with Virtual Networks over linux (VNX) tool, as new alternative for the emulation with SDN. VNX/OpenFlow approach integrates three kind of tools, an OpenFlow compliant switch (Open vSwitch), Network Operative Systems (POX, NOX and Beacon) and finally tools to control the performance and the network traffic. For the validation process, we present two VNX/OpenFlow scenarios to test the correctness of this tool. Finally, the result of this work allows the deployment of virtual scenarios with OpenFlow protocol.

Keywords— *Emulation, OpenFlow, Software Defined Networking; Virtualization.*

I. INTRODUCTION

Network data traffic has grown exponentially in the last years due the emergence of real time applications, video streaming, the rise of social networking, the introduction of cloud computing, among others. The research community has created protocols in order to cover these new needs, however the standardization process takes a long time and the improvements in communication methods and information processing are almost nonexistent [1].

Existing networks should have an open control and provide a real environment to tests with production traffic, due to these requirements the concept of Software Defined Networking arises [2]. SDN is not a new concept, rather is the result of many research projects such as the Active Networks and Ethane project [3]. SDN takes advantage of the best characteristics of these technologies (programmability, control and data plane separation), changing the way we see networks today. SDN allows the separation of data and control plane in network devices [4]. The control of the network behavior is in charge of an external device known as Network Operative System (NOS). The communication between network devices and the controller is established with a defined protocol, the most known OpenFlow [5].

Currently, a great number of enterprises like Google have incorporated OpenFlow in their infrastructures and devices, and there are some organizations, such as Open Networking Foundation (ONF), which promote the development and the widespread of OpenFlow and SDN architecture. There are few projects to test with SDN such as simulators, emulators or testbeds. One of the first OpenFlow testbed was developed by

Global Environment for Network Innovations (GENI) [6], which interconnects the principal universities of United States. Likewise, the project OpenFlow in Europe: Linking Infrastructure (OFELIA) [7] connects 8 OpenFlow islands, allowing experimentation with this technology.

Other interesting tool is ns-3 simulator [8]. Although ns-3 has support for OpenFlow, it does not work with typical controllers such as POX [9], NOX [10], Beacon [11], Floodlight [12], OpenDaylight, and so on. Instead, ns-3 has its own OpenFlow controller. Regarding OpenFlow emulators, the most known is Mininet which is used for rapidly prototyping large networks [13]. Mininet can run real applications with a great variety of topologies; however, the performance fidelity depends on the CPU capacity and the number of the emulated hosts. Additionally, there is a hybrid approach that combines simulation and emulation in one tool called EstiNet [14]. It has not problems with fidelity performance, however, it is not a free tool.

There is a wide range of tools for experimentation with virtual networks, such as the virtualization tool called Virtual Networks over linux (VNX) [15]. VNX is used in education and research fields, for instance in the experimentation with Intrusion Detection Systems (IDS), Multipath TCP (MTCP), among others. This paper presents the integration of this tool with OpenFlow protocol. For this purpose, OpenFlow-enabled switch and controllers are integrated.

This work has been divided into five sections, as follows: The second section contains an introduction of SDN and OpenFlow protocol. Then, the third section presents the description of simulation and emulation tools. Next, the fourth

section shows the VNX-OpenFlow integration process and the validation of two test scenarios. Finally, a discussion is opened in the fifth section.

II. SOFTWARE DEFINED NETWORKING

Software Defined Networking introduces a paradigm change in the network communication, facilitating the innovation and the network programmability. SDN proposes the separation between the control and the data plane in networking devices. Consequently, the network is more flexible, programmable and it has automation capabilities. The own device could carry out advanced capacities such as firewall rules, load balancing, among others.

The control of whole network is performed by a central point known as a controller. The network devices are connected with the controller through secure communication channel like Sockets Secure Layer (SSL). In the communication process is needed a standardized protocol the most known OpenFlow [5], which defines the communication rules between controller and OpenFlow compliant switches. OpenFlow offers new features that enable experimentation without expose the internal structure of switches from different vendors. For this purpose, OpenFlow delimits the basic functions of OpenFlow switches based on common characteristics of traditional Ethernet switch. OpenFlow defines three kind of tables, these are: flow, group and meter table. OpenFlow also introduces the flow concept, which can be defined as a kind of traffic such as the http requests, traffic to the same destination address, and so on. Moreover, OpenFlow establishes a pipeline in order to process the incoming packets. The packet is first matched against flow entries of flow table 0 and may continue with the next tables, depends on the result of the match in the table. Flow entries match packets based on the priority field (highest priority). If a flow entry is found, the instructions are executed (Modify packet and update match fields, update action set, update metadata). If the packet does not match with a flow entry in any table, the outcome depends on the configuration of the table miss. A possible action is to search in the next table. Based on the SDN architecture and the business requirements many tools have been developed, such as:

- Virtualization tools [16].
- Network Operating System (controllers) [9] [10].
- Virtual switches [17].
- Tools for Quality of Services and Quality of Experience [18].
- Management [19] [20].
- Optical Networks [21] [22].
- Traffic engineering and load balancing [23].
- Load Balancing [24].
- Simulation and Emulation tools [8] [13] [14].

All of these research fields are deployed and tested through some approaches; real testbeds, emulator or simulators [25]. OpenFlow testbeds [6] [7] allow the experimentation in real environments on a large scale. However, testbeds are not easily accessible by potential researchers. For its part,

simulation and emulation approaches provide facilities in terms of scalability, portability and accessibility in the case of open source tools. Nevertheless, in some cases they produce inaccurate outcomes. We describe some familiar tools ns-3, Mininet and EstiNet, as well as VNX/OpenFlow.

III. SIMULATION AND EMULATION TOOLS

NS-3 is a simulator tool focuses on research and educational fields. It is an open sources simulator that provides an extensible network platform with several external animators, data analysis and visualization tools. In order to enable the simulation with OpenFlow protocol, Ns-3 implements its OpenFlow-enabled switch and its own controller, as a modules written in C++. The switch component is known as *OpenFlowSwitchNetDevice*. This object consists of a set of net devices that represent the switch ports, according to the OpenFlow Switch Specification v0.8.9. Even though Ns-3 can be used for real-time simulations, there are some issues that the user should take into account such as the slow learning curve to use the tool, the compatibility with a basic OpenFlow version (0.89) and specially it does not run a typical OpenFlow controller. Therefore, the controller applications generated with ns-3 controller cannot be used in real network. If a controller like Pox or Floodlight was required, these will need substantial modifications.

For its part, Lantz et al. in [13] proposes Mininet, a virtualization tool for rapidly prototyping large networks in a single laptop. This tool includes OpenFlow support and combines lightweight virtualization capabilities over Linux operative system with an extensible CLI and API. A scenario built with Mininet is deployable, interactive, scalable, realistic and it can easily share. In fact, the Mininet topologies and the controller applications can be used for others researchers without modifications in the emulation environment as well as in real networks. Mininet run on virtual machine monitors like VMWare, XEN and VirtualBox or it can be installed in a Linux system. Mininet allows the deployment of hundreds of nodes, emulating OpenFlow-enabled switches, controllers like POX, virtual links and hosts. Mininet shares components like the file system, the user ID space, the kernel, device drivers, among others. Tough, Mininet is the most popular tool for SDN has limitations of performance fidelity related with the available resources, real bandwidth and the timing of the process.

A novel hybrid approach has recently presented called EstiNet [14]. This combines the best characteristics of both simulation and emulation mode in one tool. On the one hand, it allows the deployment of large networks in a flexible, easy, scalable and repeatable way. On the other hand, EstiNet takes into account the timing needs for real applications in order to obtain the same results in both, virtual and real deployments. EstiNet supports 1.3.2 OpenFlow Switch Specification and it can run NOX, POX, Floodlight, and Ryu controllers without any modifications. For this purpose, EstiNet intercepts the packets between two real applications through tunnel network interfaces and redirects the packets to the EstiNet simulation engine. The entire process is based on a simulation clock, which allows accurate results. Besides, EstiNet provides a graphical user interface for configure the scenarios and observe

the outcomes from the simulations. The results of this tool show better scalability and performance than Mininet, however their main problem is that it need a payment for the tool. The universities can embrace the EstiNet University Program. This grants a license during six months with a cost of US\$1500 or a license to 12 months for US\$2500, becoming its main disadvantage.

As we have seen, there are few tools or testbeds that allow the SDN experimentation. We present VNX a modular architecture based on plugins (Fig. 1), which allows the deployment of virtual testbeds. This tool includes the code of the previous tool VNUML [26].

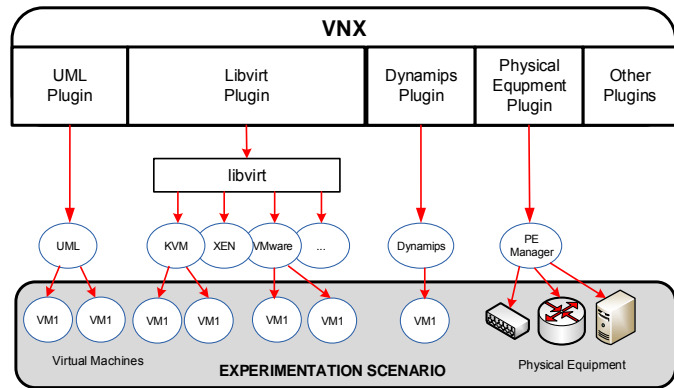


Fig. 1. VNX Architecture [15].

The plugins used by VNX are:

- UML (User Mode Linux) can be considered a hypervisor-based technique.
- libvirt allows virtualization capabilities and some virtualization platforms, such as Xen, VMware, KVM, VirtualBox, etc.
- Dynamips plugin allows the emulation the hardware of Cisco routers.
- Olive allows the integration of Juniper routers.
- Physical equipment plugin, which allows the connection between VNX physical islands.

VNX is a free tool based on Linux that allows the easy creation and management of large virtual scenarios over a single server or a cluster. The scenarios can have nodes in some physical hosts and can use different operative systems, for example Linux and Windows. In turn, each physical host can deploy their own virtual testbed. Besides, VNX allows the creation of large scenarios with hundred or even thousands of virtual machines. This process uses the copy on write technique (cow), which starting the virtual machines from a single image file known as filesystem. In this way, the nodes can share the same filesystem. The filesystem is mounted in read-only mode. If a virtual node is modified, the differences are stored in a private filesystem.

VNX is also focused on education and research. In [15] a large virtual network scenario was created. It is a laboratory for dynamic routing that involve 44 virtual devices (16 Cisco

routers, 6 Juniper routers, 6 Linux/Quagga routers, 12 end user and 4 Servers). This testbed is a typical scenario deployed with VNX and shows its potential.

One of the main SDN challenges is the integration of heterogeneous networks. VNX could provide the ideal environment to combine OpenFlow-enabled islands and legacy networks. The integration process is described in the next section.

IV. INTEGRATION AND VALIDATION

VNX should be implemented over a Linux operating system. The guidelines for configuration, modifications and filesystems are available in the official site of VNX project [27]. In order to testing with OpenFlow protocol, VNX needs the integration of some critical elements, an OpenFlow-enable switch for virtualization environments and a network operative system for network control. Additionally, it would be useful the integration of performance tools or data traffic analyzer. VNX was installed on a physical host with Ubuntu 12.04. Then, we choose two different filesystems. For controller device is desirable a graphical interface (ubuntu-12.04-gui-v024) to analyze the traffic. The second filesystem is a console interface (ubuntu-12.04-v024), which is used for simulated hosts and routers. The graphical filesystem was modified to make the controller functions, 3 of them were integrated: POX (based on Python) which is one of the most widely used today, NOX based on c++ and Python and finally Beacon which uses Java. The integration and configuration process are available in the official sites of each project. Additionally, in order to improve the functionalities of VNX/OpenFlow, three tools were installed: Wireshark, tcpdump and iperf. The wireshark tool is indispensable because originally it does not identify OpenFlow traffic. For this purpose, a dissector plugin for OpenFlow must be compiled and installed in the filesystem. Dissector allows to decode all information of specific incoming packets, in this case OpenFlow (version 1.0). Other important changes is the integration of Open vSwitch (OVS) [17]. OVS is an open source tool that allows the creation of switches in virtualization environments. OVS matches the virtual machines, providing better performance than the traditional bridge, such as VLANs, netFlow, QoS, bonding, mirroring, among others. OVS works transparently with VNX, for both legacy and OpenFlow networks. The version used in this paper is 1.4.0. After we create the .xml specification (Fig. 2).

```
<vm type="libvirt" name="C2" os="linux" subtype="kvm">
  <filesystem type="cow"/usr/share/vnx/filesystems/rootfs_ubuntu-gui/</filesystem>
  <mem>512M</mem>
  <console display="yes" id="0"/>
  <console display="no" id="1"/>
  - <if id="2" net="Net1">
    <ipv4>10.0.1.3/24</ipv4>
  </if>
  <route type="ipv4" gw="10.0.1.1">default</route>
</vm>
<host>
  - <hostif net="Net2">
    <ipv4>10.0.2.2/24</ipv4>
  </hostif>
  <route type="ipv4" gw="10.0.2.1">10.0.0.0/16</route>
</host>
```

Fig. 2. XML Specification for Design Phase.

Once we have the file with .xml specification, the virtual scenario is deployed and matched with the controller. For the validation process we replicate the topology of OpenFlow Tutorial, as a point of reference to see the VNX operation. This tutorial was developed by Stanford University [28] and it

deploys a topology (subnet 10.0.0.0/24) with 3 virtual hosts (h2, h3 and h4), an OpenFlow switch (s1) and one controller (c0). Two scenarios are presented: a basic scenario (Fig. 3a) that is identical to OpenFlow Tutorial and the second scenario incorporates more subnets and a second controller (Fig. 3b).

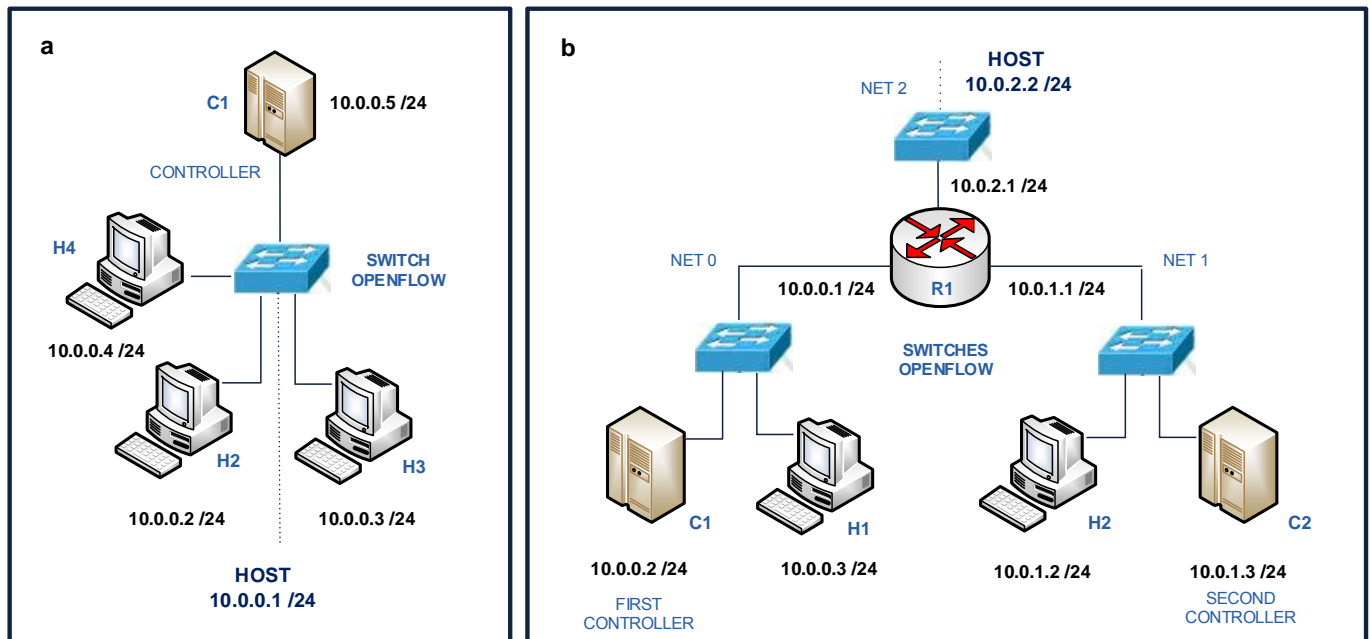


Fig. 3. (a) Scenario 1. Basic Scenario; (b) Scenario 2. Scenario with two Controllers.

The first scenario (Fig. 3a) has an OpenFlow-enabled switch and four hosts (C1, H2, H3, H4), all of them with Ubuntu 12.04. H2, H3 and H4 work with textual consoles and the controller (C1) works with a graphical console. The second scenario (Fig. 3b) is formed by five Ubuntu 12.04 virtual machines (router and hosts work with textual consoles and controllers with graphical console) according to the following structure:

- 3 switches in different subnets (Net0: 10.0.0.0/24, Net1: 10.0.1.0/24 y Net2: 10.0.2.0/24).
- 2 controllers (C1: 10.0.0.2 and C2: 10.0.1.3).
- 2 hosts (H1:10.0.0.3 and H2:10.0.1.2) each one in different subnets.
- Subnets communicate through the router (R1).

The proofs of concept of this work were made exclusively with Ubuntu virtual machines, but it is possible to use another kind of operating system. Data traffic was analyzed with Wireshark. At first, Wireshark shows only typical protocols, such as ICMP, UDP, IP, among others, because OVS works as an Ethernet switch by default.

In order to enable OpenFlow traffic, OVS must be connected with the controller. There are two configuration modes, which determine the switch behavior for a controller fail condition. These modes are:

- Fail standalone: The default configuration mode. If OVS does not receive the inactivity probe interval three times,

the OVS takes the control of the switch and it works like a normal Ethernet switch (MAC-learning switch). When the connection is lost, the switch handles the incoming packets using the OFPP_NORMAL reserved port. Moreover, the switch will attempt to connect with the controller. These mode is usually available in OpenFlow hybrid switches.

- Fail secure: In this mode the OVS cannot take the network control if the controller fails. The network will be uncommunicated during the failure. Then, OVS will attempt to connect with the controller, until obtain a response. This mode is commonly used to avoid forwarding loops.

Once the communication is established, the controller (or controllers) should maintain the links with all switches. There are three kinds of roles for the connection. The default role is OFPCR_ROLE_EQUAL and it allows full control over the network. The second role is known as OFPCR_ROLE_SLAVE, in which switches are configured in read only mode, therefore the controller has limited control. The third role, OFPCR_ROLE_MASTER works in the same way that OFPCR_ROLE_EQUAL, but there is only one controller with this role, other controllers are changed to slave role. In the second scenario all switches are connected with C1 and C2 controllers in EQUAL role. In this way, we provide redundancy to the second scenario.

Proofs were made with standalone and secure mode in both scenarios. We used POX controller with three applications,

forwarding.l2_learning, forwarding.l3_learning and forwarding.hub. Additionally, we wrote scripts in order to automate the process. These scripts contain the code for the deployment of the above mentioned scenarios and the establishment of links between switches and controller.

In both scenarios data traffic was generated with ICMP and web requests between the hosts of the topologies. OFP (message for the establishment of network communication), OFP-ARP, OFP-ICMP (packet-in, packet-out) messages were captured with Wireshark analyzer and tcpdump tools as shown in Fig. 4.

26	4.088734	10.0.0.2	10.0.2.2	OFP	74 Echo Reply
27	4.088827	10.0.0.2	10.0.2.2	OFP	74 Echo Reply
28	4.101774	02:fd:00:00:03:01	02:fd:00:00:01:01	OFP+ARP	126 Packet In
29	4.105569	10.0.0.2	10.0.2.2	OFP	90 Packet Out
32	4.106380	02:fd:00:00:01:01	02:fd:00:00:03:01	OFP+ARP	126 Packet In
33	4.108450	10.0.0.2	10.0.2.2	OFP	90 Packet Out
37	4.425297	10.0.0.3	10.0.1.2	OFP+ICMP	182 Packet In
39	5.038602	10.0.1.2	10.0.0.3	OFP+ICMP	182 Packet In
41	5.425545	10.0.0.3	10.0.1.2	OFP+ICMP	182 Packet In

Fig. 4. Traffic Capture Scenario 2.

Fig. 4 shows an ICMP proof from host h1 (10.0.0.3) to host h2 (10.0.1.2) performed in the second scenario, with the component forwarding.l2_learning of POX controller and in standalone mode.

Both scenarios work properly with OpenFlow protocol, however in second scenario there were duplicated messages (from controllers C1 and C2). This is because OpenFlow does not define coordination mechanisms among controllers in the same network or in different domains [29]. At present, this process is done with other components. For instance, Fonseca et al. in [30] introduces the CPRcovery component, which allows keeping the consistency between the primary and backup controllers. This component provides seamless transition between the primary and secondary controller through two steps, the replication phase (maintain updated data) and the recovery phase. The replication phase acts during the normal network behavior and the recovery phase acts in case of failure. Another challenge in large topologies is the communication among controllers in different SDN domains. At the present time, Internet Engineering Task Force (IETF) is working in a standard called interfacing SDN Domain Controllers (SDNi) for exchange routing information (network topology views, network conditions, event reports) and application requirement.

A general overview for the whole process in order to interact with VNX/OpenFlow scenarios is shown in figure 5. The first phase consist of the design and creation of VNX scenarios based on .xml specification. The second phase is related to the deployment or destruction of these scenarios through specific commands (vnx -f -v --create). Then, the

controller must be connected with the switches and the user should configure the operation mode (standalone, secure, equal, slave, master). The user can create their own topologies and programs with the controller and finally can interact with the OpenFlow testbed.

V. CONCLUSION AND DISCUSSION

This work presents the integration process between VNX tool and OpenFlow protocol. The filesystems used by virtual machines and nodes was modified. We create a SDN environment through the integration of two main components: an OpenFlow compliant switch (Open vSwitch) and three network operating systems (NOX, POX and Beacon). Besides the controller has incorporated some performance and analyzer tools, these are Wireshark, tcpdump and iperf. Proofs of concept were carried out with POX components and two configuration modes (secure and standalone).

We can verified the exchange of OpenFlow messages (OFP+ARP, OFP+ICMP Packet In, OFP packet Out) with Wireshark analyzer. Although in the validation process we only used Ubuntu, future proofs can use multiple operating systems such as Windows. Now the user can create their own topologies and controller programs in order to experiment with OpenFlow protocol and SDN technology, which was the main objective of this work.

Today, VNX allows the deployment of large and complex OpenFlow networks in distributed environments. VNX allows not only the deployment of virtual scenarios in a single laptop, but also allows the inclusion of physical equipment (each one can have its own scenario with virtual machines), that is, VNX works in distributed scenarios. This is the main contribution of VNX over Mininet, since the communication between two scenarios in Mininet is a complex process. In this way, VNX enable the communication between OpenFlow networks and legacy networks that is one of the main challenges of SDN, the transition and migration process between heterogeneous networks. Besides, take into account that virtual scenarios may include Cisco and Juniper devices, therefore inside the virtual scenarios we could test with OpenFlow and no OpenFlow networks. Moreover, VNX allows the easy experimentation with specific services such as multimedia applications, deployment of servers, among others. The developer can customize the filesystem of the hosts and in this way, testing their new ideas and applications.

VNX also allows another kinds of operating systems for the virtual machines, such as Debian, Windows and Fedora. This is another strong point compared with Mininet, which uses only a Linux kernel. If a user want to test a Windows application over an OpenFlow network, the windows filesystem may include the application code.

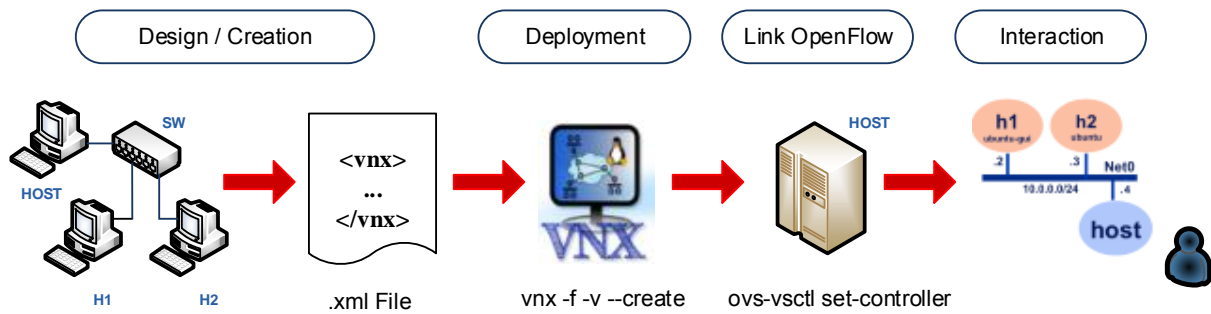


Fig. 5. Workflow of VNX/OpenFlow.

ACKNOWLEDGMENT

The research leading to these results has been partially funded by the European Union's H2020 Program under the project SELFNET (671672). Lorena Isabel Barona López and Ángel Leonardo Valdivieso Caraguay are supported by the Secretaría Nacional de Educación Superior, Ciencia, Tecnología e Innovación SENESCYT (Quito, Ecuador) under Convocatoria Abierta 2012 and 2013 Scholarship Program. This work was partially supported by the "Programa de Financiación de Grupos de Investigación UCM validados de la Universidad Complutense de Madrid – Banco Santander".

The authors would like to thank to David Fernández Cambronero for his comments and suggestions about VNX tool and Ana Lucila Sandoval Orozco for her valuable comments and suggestions to improve the quality of the paper.

REFERENCES

- [1] W. Stallings, "Software Defined Networks and OpenFlow," in *The Internet Protocol Journal*, vol. 16, no. 1, March 2013, pp. 2-14.
- [2] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, J. Turner, "OpenFlow: Enabling Innovation in Campus Networks," in *ACM SIGCOMM Computer Communication Review*, vol. 38, no. 2, April 2008, pp. 69-74.
- [3] M. Casado, M. J. Freedman, J. Pettit, J. Luo, N. McKeown, S. Shenker, "Ethane: Taking Control of the Enterprise," in *Proceedings of the ACM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, New York, USA, August 2007, pp. 1-12.
- [4] A. L. Valdivieso Caraguay, L. I. Barona López, L. J. García Villalba, "Evolution and Challenges of Software Defined Networking," in *Proceedings of the Workshop on Software Defined Networks for Future Networks and Services*, Trento, Italy, November 2013, pp. 47-55.
- [5] O. S. Consortium et al., "OpenFlow Switch Specification v.1.3.4," March 2014 pp. 1-171.
- [6] C. Elliott, "GENI: Opening Up New Classes of Experiments in Global Networking," in *IEEE Internet Computing*, vol. 1, January 2010, pp. 39-42.
- [7] M. Suñé, L. Bergesio, H. Woesner, T. Rothe, A. Köpsel, D. Colle, B. Puype, D. Simeonidou, R. Nejabati, M. Channegowda, M. Kind, T. Dietz, A. Autenrieth, V. Kotronis, E. Salvadori, S. Salsano, M. Körner, S. Sharma, "Design and implementation of the OFELIA FP7 facility: The European OpenFlow testbed," in *Computer Networks*, vol. 61, March 2014, pp. 132-150.
- [8] T. R. Henderson, M. Lacey, G. F. Riley, "Network Simulations with the ns-3 Simulator," in *Proceedings of the ACM SIGCOMM'08*, Seattle, WA, USA, August 2008, pp.17-22.
- [9] POX, <https://openflow.stanford.edu/display/ONL/POX+Wiki>.
- [10] N. Gude, T. Koponen, J. Pettit, B. Pfaff, M. Casado, N. McKeown, S. Shenker, "NOX: Towards an Operating System for Networks," in *ACM SIGCOMM Computer Communication Review*, vol. 38, no. 3, July 2008, pp. 105-110.
- [11] D. Erickson, "The Beacon Openflow Controller," in *Proceedings of the Second ACM SIGCOMM Workshop on Hot Topics in Software Defined Networking*, New York, NY, USA, August 2013, pp. 13-18.
- [12] Floodlight project, <http://www.projectfloodlight.org/>.
- [13] B. Lantz, B. Heller, N. McKeown, "A Network in a Laptop: Rapid Prototyping for Software-Defined Networks," in *Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks*, New York, NY, USA, October 2010, pp. 1-6.
- [14] S. Y. Wang, C. L. Chou, C. M. Yang, "EstiNet OpenFlow Network Simulator and Emulator," in *IEEE Communications Magazine*, vol. 51, no. 9, September 2013, pp. 110-117.
- [15] D. Fernández, A. Cordero, J. Somavilla, J. Rodriguez, A. Corchero, L. Tarrafeta, F. Galán, "Distributed Virtual Scenarios over multi-Host Linux Environments," in *Proceedings of the 5th IEEE International DMTF Academic Alliance Workshop on Systems and Virtualization Management*, Paris, France, October 2011, pp. 1-8.
- [16] R. Sherwood, G. Gibb, K.-K. Yap, G. Appenzeller, M. Casado, N. McKeown, G. Parulkar, "Flowvisor: A Network Virtualization Layer," in *Technical Report OpenFlow Switch Consortium*, October 2009, pp. 1-15.
- [17] B. Pfaff, J. Pettit, K. Amidon, M. Casado, T. Koponen, S. Shenker, "Extending Networking into the Virtualization Layer," in *Proceedings of the Eight ACM Workshop on Hot Topics in Networks*, HotNets-VIII, HOTNETS '09, New York City, NY, USA, October 2009, pp. 1-6.
- [18] A. Kessler, L. Skorin-Kapov, O. Dobrijevic, M. Matijasevic, P. Dely, "Towards QoE-driven Multimedia Service Negotiation and Path Optimization with Software Defined Networking," in *Proceedings of the 20th IEEE International Conference on Software, Telecommunications and Computer Networks*, (SoftCOM), Split, Croatia, vol. 1, September 2012, pp. 1-5.
- [19] R. Benesby, P. Fonseca, E. Mota, A. Passito, "An Inter-AS Routing Component for Software-Defined Networks," in *Proceedings of the IEEE Network Operations and Management Symposium*, Maui, Hawaii, USA, April 2012, pp. 138-145.
- [20] F. Farias, J. Salvatti, E. Cerqueira, A. Abelem, "A Proposal Management of the Legacy Network Environment Using Openflow Control Plane," in *Proceedings of the IEEE Network Operations and Management Symposium*, Maui, USA, April 2012, pp. 1143-1150.
- [21] S. Das, G. Parulkar, N. McKeown, P. Singh, D. Getachew, L. Ong, "Packet and Circuit Network Convergence with OpenFlow," in *Proceedings of the IEEE Conference on Optical Fiber Communication (OFC), collocated National Fiber Optic Engineers Conference (OFC/NFOEC)*, San Diego, CA, USA, March 2010, pp. 1-3.
- [22] M. Channegowda, R. Nejabati, M. Rashidi Fard, S. Peng, N. Amaya, G. Zervas, D. Simeonidou, R. Vilalta, R. Casellas, R. Martínez, "First Demonstration of an OpenFlow based Software-Defined Optical

- Network Employing Packet, Fixed and Flexible DWDM Grid Technologies on an International Multi-Domain Testbed,” in *Proceedings of the European Conference and Exhibition on Optical Communication*, Amsterdam, Netherlands, September 2012, pp. 1-3.
- [23] N. Handigol, S. Seetharaman, M. Flajslik, N. McKeown, R. Johari, “Plug-n-Serve: Load-Balancing Web Traffic using OpenFlow,” in *Proceedings of ACM SIGCOMM Demo*, Barcelona, Spain, August 2009, pp. 1-2.
- [24] I. F. Akyildiz, A. Lee, P. Wang, M. Luo, W. Chou. “A Roadmap for Traffic Engineering in SDN-OpenFlow Networks”. in *Computer Networks*, vol.71, June 2014, pp. 1-30.
- [25] M. Kobayashi, S. Seetharaman, G. Parulkar, G. Appenzeller, J. Little, J. Reijndam, P. Weissmann, N. McKeown, “Maturing of OpenFlow and Software-defined Networking through Deployments,” in *Computer Networks*, vol. 61, March 2014, pp. 151-175.
- [26] F. Galán, D. Fernández, W. Fuertes, M. Gómez, J. E. L. de Vergara, “Scenario-based Virtual Network Infrastructure Management in Research and Educational Testbeds with VNUML,” in *Annals of Telecommunications-Annales des Telecommunications*, vol. 64, May 2009, pp. 305- 323.
- [27] “Virtual Networks over linuX (VNX),” http://web.dit.upm.es/vnxwiki/index.php/Main_Page.
- [28] OpenFlow Tutorial, http://www.openflow.org/wk/index.php/OpenFlow_Tutorial.
- [29] A. L. Valdivieso Caraguay, A. Benito Peral, L. I. Barona López, L. J. García Villalba, “SDN: Evolution and Opportunities in the Development IoT Applications,” in *International Journal of Distributed Sensor Networks*, vol. 2014, May 2014 pp. 1-10.
- [30] P. Fonseca, R. Bennesby, E. Mota, A. Passito, “A Replication Component for Resilient OpenFlow-based Networking,” in *Proceedings of the IEEE Network Operations and Management Symposium, Maui, Hawaii, USA*, April 2012, pp. 933-939.

An Overview of Integration of Mobile Infrastructure with SDN/NFV Networks

Ángel Leonardo Valdivieso Caraguay, Lorena Isabel Barona López, Luis Javier García Villalba

Group of Analysis, Security and Systems (GASS)
Department of Software Engineering and Artificial Intelligence (DISIA)
Faculty of Information Technology and Computer Science, Office 431
Universidad Complutense de Madrid (UCM)
Calle Profesor José García Santesmases, 9
Ciudad Universitaria, 28040 Madrid, Spain
Email: {angevald, lorebaro}@ucm.es, javiergv@fdi.ucm.es

Abstract— The growing number of on-line applications and services running on wireless and mobile devices has been limited by the rigidity of actual IT infrastructure, in which the closed union between data and control planes limits the possibility of customize the network behavior. In this context, the concepts of SDN and NFV appear as a viable solution to open the infrastructure to developers in order to create new services and applications. In this work, we describe the concepts of SDN, NFV and analyze the possibility of integrate these technologies in mobile networks. Furthermore, we present the last projects and an architecture proposal focused on this direction. Finally, we discuss the trends and challenges in order to implement these advances in production networks.

Keywords—*Mobile Network; Network Function Virtualization; OpenFlow; Software Defined Networking.*

I. INTRODUCTION

The diversity of network infrastructures has enabled the increase of connectivity among users and consequently it has promoted the establishment of new business models. This new digital environment requires an IT infrastructure capable to ensure high level of Quality of Service (QoS) and customization of applications. However, the heterogeneity of cellular and wireless technologies and the current configuration techniques complicate the control and management of the network.

The IT infrastructure is composed by a set of hardware devices running proprietary software that analyzes the traffic and selects the optimal route to the destination. In this scenario, the network administrator does not have access to modify the internal operation of the device. Instead, the administrator can only configure a minimum set of parameters to modify the network behavior. Moreover, the inclusion of new services requires the individual updating of devices or the complete replacement of hardware infrastructure. For this reasons, the idea of separate the data plane and control plane in order to customize the network behavior has gained importance. Similarly, the possibility of encapsulate the different network functions based on actual network conditions can optimize the allocation of available resources.

The concepts of Software Defined Networking (SDN) and Network Function Virtualization (NFV) have changed the vision of typical network infrastructure. SDN separates the data and control planes in network devices and establishes a centralized control of the network behavior. This architecture

enables to the network administrator the possibility to design and develop “network applications” and dynamically control the network. For its part, NFV allows the deployment of virtualized network functions (e.g. load balancers, firewalls) as virtual instances over standardized hardware (storage, network and servers). This technology integrates the use of different resources (servers, storage, IT-hardware), enhances the scalability of the network services and reduces the capital and operational cost.

A techno-economic analysis in mobile infrastructure reveals that the benefits of the introduction of SDN and virtualization techniques could decrease the capital expenditures. The capital expenditure could be reduced around 13.81 % in a SDN scenario [1]. It is clear that architectures based on SDN offer multiple potential advantages for telecom operators [2], for instance, the possibility of deploy Radio Base Stations in the Cloud [3] or integrating LTE network elements with SDN switches managed from the cloud [4]. In this context, the industry and research community go a step further in this direction and have been combining their efforts in multiple projects such as OpenRoads [5], SoftCell [6] as well as European Projects such as T-NOVA [7], UNIFY [8] among others. In this piece of work, we describe the concept and the evolution of SDN and NFV in the last years. Furthermore, the integration of mobile infrastructure with SDN/NFV as well as the trend and challenges to implement these technologies in production networks are analyzed.

The work is structured as follows: in Chapter II the concepts of SDN and NFV are presented. Next, Chapter III reviews the integration of mobile networks with both

technologies. Chapter IV analyzes the trends and challenges and provides an initial SD/NFV architecture. Finally, Chapter V presents the conclusions.

II. SDN/NFV

A typical network device is composed by an integrated data plane and control plane. The data plane receives the packet, reads the header information, sends the information to the control plane and forwards the packet to the next network device. For its part, the control plane analyzes the information provided by the data plane and executes a routing algorithm to establish the optimal route to the destination. Once the route is chosen, the control plane sends the decision to the data plane. However, the limited coordination and access to the configuration of the devices (closed technology) has limited the development of customized network applications and QoS services.

Software Defined Networking is a new network paradigm that removes the rigidity present on current architectures and improves flexibility and management in networks. SDN decouples the control plane and the data plane in network devices and establish an open communication interface between them. In addition, SDN proposes a centralized control of the network and open APIs to facilitate the development of high level network applications and services. OpenFlow is the first SDN standard that has been widely used in different research projects [9] [10]. OpenFlow is designed based on the actual flow tables located in traditional network devices and opens those up. The controller uses the OpenFlow protocol [11] to configure the flow tables in switches. Figure 1 shows the differences between SDN and traditional architectures.

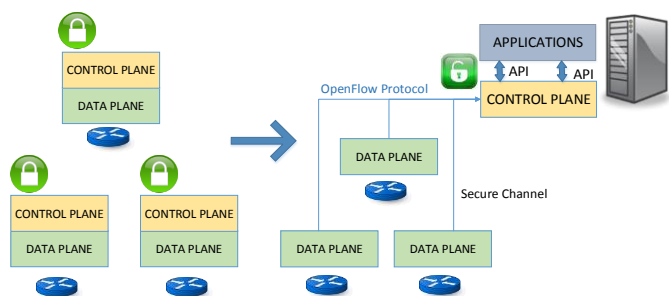


Fig. 1. Comparison between traditional and SDN architectures [10]

Another limitation of the actual infrastructure is the highly amount of network devices, each operating their own private software and highly dependent in proprietary hardware. For this reason, the design and installation of new services usually require the individual software updating or the replacement of hardware. This rigidity increases the installation and operational costs. In this context, the Network Function Virtualization NFV concept has gained in importance in the telecommunications industry.

Network Function Virtualization proposes the transferring of the different network functions (routing, firewall, deep packet inspection DPI, gateway) as virtual software-based

applications executed in IT platforms (servers, switches and storage). This new vision of IT services provides a major flexibility and scalability, facilitates the development cycles and reduce costs. Figure 2 describes the differences between NFV and traditional architectures.

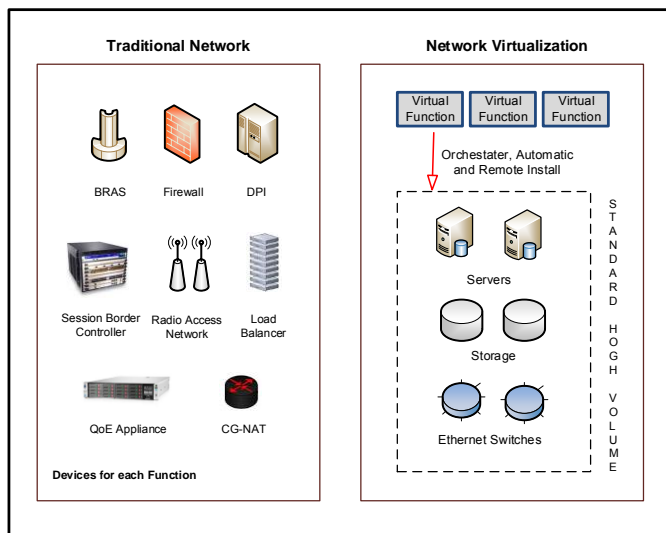


Fig. 2. Comparison between traditional and NFV paradigms

The NFV architecture identifies three principal modules: the Network Function Virtualization Infrastructure (NFVI) that includes all hardware resources, the Virtualized Network Function (VNF) that represents a network functions executed over the NFVI and the NFV Management and Orchestrator (NFV M&O) responsible for coordinate the execution of the different network functions (NF) over the infrastructure [12].

It is important to note that SDN and NFV concepts are different but complementary. Thanks to the SDN separation of data and control planes in network devices, the integration of NFV in virtualized IT environments is feasible. However, the implementation in production networks has several challenges to be addressed [13].

III. INTEGRATION OF WIRELESS NETWORKS WITH SDN/NFV

In the last years, SDN approach has been expanded to mobile networks, giving rise the concept of Software Defined Wireless Networks [5] [14] [15] [16] or Wireless Mesh Software Defined Networks [17] [18]. Similarly, the integration between SDN and mobile technologies (LTE) has gained the attention of industry and research community [19] [20] [21] as well as the close relationship between SDN and NFV. For instance, EmPOWER [22] shows a testbed composed of 30 nodes that facilitates the deployment of SDN/NFV experiments for WIFI networks and it also provides monitoring tools in order to control the energy consumption. In the wireless field, some research intends to apply OpenFlow in order to enhance these types of networks.

OpenRoads [5] presents an architecture of three layers (flow, slicing and controller) based on OpenFlow protocol and SNMP in order to innovate in WIMAX and WIFI networks. In the same way as OpenFlow wired networks, the wireless devices (Access Point or Wimax base station) have a flow table which are controlled through the controller. For its part, the slicing layer divides the data traffic through the FlowVisor tool [23] and NOX controller is the brain of the network control. The deployment contains 85 Access Points and two Wimax Base stations (over Stanford campus network) and provides functions such as hard handover, bicasting, Hoolock, among others.

Dely et al. in [18] presents an approach to improve the mobility in Wireless Mesh Networks (WMN). For this purpose, it introduces a mesh router known as a Mesh Access Point (MAP), which forwarding the traffic to the destination through other mesh routers or gateways. It is important to note that MAPs have OpenFlow support. Each node has some physical wireless cards and these in turn are divided into two virtual interfaces, one related with the data plane and the other for control traffic. The data interface is related with OpenFlow datapath. The gateways allow connectivity with the outside and each mesh router has an agent in order to monitoring the links, channel utilization, and others. Moreover the core network has two elements: the Monitoring and Control Server (MCS) and NOX controller. MCS builds a topology database with the information from mesh routers and NOX controls the mesh network. Proofs of concept were conducted over an experimental Wireless Broadband Mesh Network (KAUMesh), which is based on 802.11a/b/g standard. These tests were focused on mobility capacities, when clients move rapidly between different MAPs.

Regarding to cellular networks, the advances are limited and not homogeneous due each research applies different approaches and focuses on diverse elements, the Radio Access network (RAN) and the Evolved Packet Core (EPC).

On one hand, SoftRAN [24] framework uses the SDN concept in order to improve the RAN performance. SoftRAN has a whole view of interference and load of each node and in this way coordinates the allocation of radio resources, especially in dense networks. Each base station sends periodically information to controller and it is saved in a data base, which contains the following elements: an interference map, the flow records and the network operator preferences. SoftRAN was tested with some use cases such as load balancing and utility optimization.

On the other hand, there are some approaches focused on the core part of cellular networks. CellSDN [25] [26] provides an architecture with advances characteristics, such as the slicing of the network resources, better packet classification through deep packet inspection functionalities, scalability via local switch agents and the creation of applications based on the user attributes (network provider, device type).

For its parts, SoftCell [6] enhances the scalability and flexibility in SDN/LTE networks through the analysis of workload and the implementation of fine-grained policies.

Softcell also aggregates the traffic based on different aspects such as the base station, mobile devices and the service polices. Each base station is connected with an access switch. This switch has OpenFlow support and is supervised by the controller.

Similarly, MobileFlow [21] takes advantage from SDN and data center concepts to enable and foster the innovation in carrier networks. The main elements of the architecture are MobileFlow Forwarding Engine (MMFE) and MobileFlow Controller (MFC). MMFE has support to mobile network tunnel capacities and allows the integration with legacy EPC equipment. For this reason, the MMFE is considered the data plane. Each MMFE is controlled by the MFC (control plane). The implementation and validation process consists on a prototype based on x86 servers and OpenFlow components.

Furthermore, some projects could be applied to SDN/mobile networks (Wireless and cellular). These projects not only take into account SDN technology but also another key enabler technologies such as NFV, cloud computing, advances virtualization techniques, among others.

For instance, T-NOVA project [7] aims the design and implementation of a framework to allow operators the deployment of virtualized Network Functions (NF) over Network/IT infrastructures. This virtual network appliances are developed in software using SDN/NFV and eliminate the need of acquire, install and maintain specialized hardware. The framework will enable an open API for developers to the design and develop of NF appliances.

UNIFY (Unifying Cloud and Carrier Networks) [8] considers the entire network (home networks to data centers) as a “unified production environment”, focusing on telco functions. UNIFY combines the benefits of cloud computing and virtualization in order to build a new architecture that optimizes data traffic flows and allows the dynamic placement of networking, computer and storage components. The consortium creates a model with advanced programmability, new languages, algorithms and management tools to optimize data traffic across networks. UNIFY intends to design a universal hardware node in order to support network functions and traditional data center workloads. The whole architecture allows agility (velocity), simplicity (automation), flexibility (granularity) and programmability of the services, providing and open environment for the deployment of these services and, at the same time, reducing the costs. UNIFY will derive a framework which supports a variety of services such as OpenFlow, Network Function Virtualization, and so on. Besides, this project is focused on three areas. First, infrastructure virtualization, second, flexible service chaining and thirdly, network service chain invocation (programmability interfaces).

CROWD project (Connectivity management for eneRgy Optimised Wireless Dense networks) [27] proposes a novel architecture in order to enhance very dense and heterogeneous wireless networks (Dense Nets). CROWD promotes a paradigm change in this kind of networks through global network cooperation, fine and dynamic network configuration,

resources on demand, among others. For this purpose, this project uses SDN and OpenFlow protocol as an enabler concepts to control and manage in an efficient way the resources of Dense Nets. CROWD architecture has two kind of controllers: local and regional. The infrastructure layer consists of base stations (eNBs or Wifi AP) which are configurable via OpenFlow. CROWD provides dynamic controller placement, dynamic backhaul reconfiguration, energy optimization, MAC optimization mechanism and ensures user quality of experience.

CITYFLOW project (OpenFlow City Experiment – Linking Infrastructure and Applications) [28] introduces the use of virtual path slice (VPS) technology at large scale on an OpenFlow network. This project emulates a city with one million inhabitants, with OpenFlow support and taking into account network topologies over xDSL, LTE and Fiber technologies.

Moreover, there are some facilities that allow the experimentation with SDN in wireless environment, such as OFELIA (Open Flow in Europe: Linking Infrastructure and Applications) [29] and the above mentioned deployments, OpenRoads [5] and KAUMesh [18]. OFELIA provides an environment to investigate and validate revolutionary ideas. OFELIA has a set on ten islands over Europe based on OpenFlow technology. Likewise, SmartFIRE [30] develops a large-scale testbed located in South Korea and Europe. The European zone is composed by three different and heterogeneous islands. Two islands belong to the OFELIA project (iMinds and UMU) and the other is part of the OpenLab federation testbed (UTH testbed). For its part, South Korea testbed includes OpenFlow islands in different institutes, for example Electronics and Telecommunications Research Institute (ETRI) and Seoul National University (SNU).

All these advancements are in an early stage but the initial results are promising. Next, we present the current trends and challenges and a possible architecture aligned with the SDN and NFV concepts.

IV. TRENDS AND CHALLENGES

Nowadays, the variety of mobile networks providing different services and applications requires a mobile infrastructure capable of provide high levels of security, performance and QoS. This means that current mobile networks requires a standardized environment, wherein foster the innovation and introduction of new services would be possible in less time and with the lowest investment. This may be achieved through the synergy of NFV and SDN. On one hand, SDN enhances the control and management of network devices through the centralized control. On the other hand, NFV reduces the investment by means of sharing resources not only physical infrastructures but also network functions. This means the reduction in capital (Capex) and operational costs (Opex) that is the main limitation of carrier and service providers. In the context of mobile networks, there are some challenges that a

SDN/NFV approach may solve. Next, we describe the actual issues and trends.

Rapid innovation: The combination of SDN and NFV reduces the time to market of new services, through the resource virtualization and centralized control in different locations over an standardized environment. This eliminates the vendor dependence and increases the benefits for stakeholders.

Mobile traffic monitoring and management: SDN allows fine-grained control of the network data traffic and resources. This is especially important for handover, where OpenFlow may facilitate the change between nodes. Additionally, the traffic could be classified and managed based on the kind of flow, aggregation criteria (cell, user equipment, etc), flow rate, occupation of the resources (channels, links, base stations or AP, available bandwidth), among others. For instance, could be possible connect users to multiple networks or defines threshold parameters (bandwidth, location), allowing the easy change between them. Other applications may include the dynamic resources management of wireless backhaul or the capacity aggregation not only with one technology but also combining different technologies.

Energy efficiency: The traffic load is changing constantly according different factors, time, location, special events, among others. On one hand, SDN may enable the optimization of the power consumption based on real time conditions. In this way, the SDN controller could increase or decrease the number of resources allocated. On the other hand, NFV may decrease the number of devices due its flexibility and sharing capabilities.

Scalability and flexibility: Nowadays, the introduction or extension of new services is not easy because current architectures are closed. Its process requires a long time or in some case is not performed due the investment is greater than the economy benefits SDN/NFV facilitates the service scalability and allows the reutilization of infrastructures and applications. Moreover, mobile networks are more flexible because SDN/NFV approach is aware of network conditions and changing traffic patterns.

Sharing infrastructure or services: SDN monitoring and VNF virtualization capabilities enable to share infrastructures and network resources. A service provider (SP) could deploy their network functions in the infrastructure of another SP, or use the applications (of another SP) in their own infrastructure. All of these activities are managed by the SDN controller. As a result of this, SDN/NFV introduces new capabilities in billing services. This generates more revenue for stakeholders, the first SP obtains revenue from the service and the second with the infrastructure lease. However, this is an ideal environment to network business; there are some legacy concerns that would be solved or agreed before this scheme can perform.

Inter-Cell Interference: Several APs or base stations in the same location could produce interference each other due the cell overlapping, bad coordination of subcarriers, among others. Consequently, it produces degradation of quality of services (QoS). In this context, SDN enables the easy

management of radio resources by means the centralized control and the global view of the network.

Security: The full picture of network events of SDN allows a better control and the detection of anomalous activities. The controller could provide pieces of software that acts like Intrusion Detection System (IDS), firewalls or another security function.

The advances in these concerns are in preliminary state and require the effort and coordination of vendors, researchers and the organism working in these areas, such as Open Networking Foundation (ONF), International Telecommunication Union (ITU) or the European Telecommunications Standards Institute (ETSI), among others [31].

The Wireless and Mobile Working Group (WMWG) aids to promote and extend the ONF approaches in this field, it includes the incorporation of OpenFlow protocol with mobile networks, following the current standards such as 3GPP, IEEE and others. ITU tries to standardize SDN for telecom carriers. For instance, Joint Coordination Activity on SDN (JCA-SDN) coordinates the ideas from different Standard Developing Organizations (SDO) and open sources activities. Other group (SG11) is discussing SDN signaling. For its part, Internet Research Task Force (IRTF) has created the Software-Defined Networking Research Group (SDNRG) and Network Function Virtualization Research Group (NFVRG), which analyze the approaches that can be used in both technologies. Moreover, ETSI-ISG has delivered some initial requirements, service models and use cases for Network Function Virtualization.

We have presented the benefits of SDN/NFV in mobile networks. Based on the premises of both technologies, a whole overview of a possible framework is shown in Figure 3.

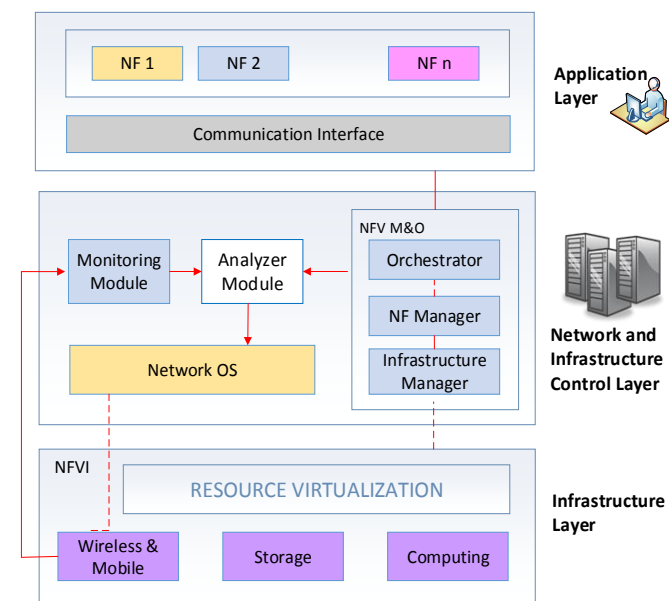


Fig. 3. SDN/NFV Architecture.

This architecture takes into account SDN and NFV technologies. On one hand, the framework presents a layered structure: data, control and application layers, in the same way that SDN architecture. Moreover, it takes advantage of NFV concept to allow the easy implementation and management of network functions, without the need to increase the hardware devices.

In the data layer, we have the current mobile infrastructure of the network operator providing support to a wide range of wireless and cellular technologies like WiFi, LTE, UTMS, GSM, among others. On top of this hardware layer, there is a virtualization layer to enable the virtualization of hardware devices. The resources could be in different locations and data centers and takes into account three components:

- **Networking:** These devices incorporate mobile technologies and OpenFlow protocol.
- **Storage:** This element can include Object storage or block storage (Swift and Cinder OpenStack) or another novel techniques.
- **Computing:** It include high volume servers. It could also use Openstack Nova.

The control layer is in charge of monitoring, analysis, management and orchestration of devices. Consist of four modules: monitoring, analyzer, network OS and NFV M&O.

- **Monitoring Module:** This module is able to provide the complete low-level overview of the managed systems by mean of gathering metrics coming from different network devices.
- **Analyzer Module:** This module could give a deep analysis of the data in order to determine the suited behavior of the network. This module also can infer the recommended behavior of the network. The techniques used in the analysis can include: data mining, learning algorithms, pattern recognition, among others.
- **Network OS:** This module control de basic functions of the control layer. Also, it uses the OpenFlow or similar protocols to send instruction to the Infrastructure Layer elements. Its functionality is similar with an Operating Systems OS in computing.
- **NFV M&O:** This module determines and organizes the actions to be executed in the system, the orchestration, the management of the resources and the control functions.

On the top of the architecture is located the application layer, which consists of two basic modules:

- **Communication Interface:** This module enables an open API to programmers to facilitate the development of new services.
- **Network Functions:** This module presents an scalable structure to create customized network functions or control applications.

This architecture enables users and developers a global view of IT infrastructure. Furthermore, the elements located on

SDN/NFV control layer can adapt the network resources depending on the actual situation of the network and dynamically respond to failures or degradation of network performance.

V. CONCLUSION

The integration of digital services distributed over multiple mobile devices (laptop, tablet, cell phone, IoT) sharing high amounts of data (VoIP, streaming, digital images, e-gaming) have been limited by the closed-access and rigidity of actual IT infrastructure. Software Defined Networking and Network Function Virtualization have emerged as a part of the solution for the openness of the infrastructure and enabling to network administrator the dynamically customization of the network behavior.

This work presents a whole overview of the limitations of current IT infrastructure and introduces the novel concepts of SDN and NFV. Similarly, we describe the recent projects based on the integration of SDN/NFV with mobile infrastructure. The current trends and challenges in order to implement these advances in production networks are analyzed. Finally, we present an SDN/NFV architecture that integrates mobile and wireless technologies. It is clear that these paradigms bring new opportunities and create new business models for users, operators and service providers. However, it is fundamental the coordination between research community, industry and service operators in order to implement these advances in production networks.

ACKNOWLEDGMENT

The research leading to these results has been partially funded by the European Union's H2020 Program under the project SELFNET (671672). Ángel Leonardo Valdivieso Caraguay and Lorena Isabel Barona López are supported by the Secretaría Nacional de Educación Superior, Ciencia, Tecnología e Innovación SENESCYT (Quito, Ecuador) under Convocatoria Abierta 2012 and 2013 Scholarship Program. This work was partially supported by the "Programa de Financiación de Grupos de Investigación UCM validados de la Universidad Complutense de Madrid – Banco Santander".

REFERENCES

- [1] B. Naudts, M. Kind, F. Westphal, S. Verbrugge, D. Colle, M. Pickavet. "Techno-economic Analysis of Software Defined Networking as Architecture for the Virtualization of a Mobile Network," in *Proceedings of the European Workshop on Software Defined Networking*, EWSDN, Darmstadt, Germany, October 2012, pp. 67-72.
- [2] J. Q. Wang, F. Haijing, C. Chang. "Software Defined Networking for Telecom Operators: Architecture and Applications," in *Proceedings of the 8th International Conference on Communications and Networking in China*, CHINACOM, Guilin, China, August 2013, pp. 828-833.
- [3] B. Haberland, F. Derakhshan, H. Grob-Lipski, R. Klotsche, W. Rehm, P. Scheffczyk, M. Soellner. "Radio Base Stations in the Cloud," in *Bell Labs Technical Journal*, vol. 18, no. 1, June 2013, pp 129-152.
- [4] J. Costa-Requena. "SDN Integration in LTE Mobile Backhaul Networks," in *Proceedings of the International Conference on Information Networking*, ICOIN, Phuket, Thailand, February 2014, pp. 264-269.
- [5] K. K. Yap, M. Kobayashi, R. Sherwood, T. Y. Huang, M. Chan, N. Handigol, N. McKeown. "OpenRoads: Empowering Research in Mobile Networks," in *ACM SIGCOMM Computer Communication Review*, vol. 40, no. 1, January 2010, pp. 125-126.
- [6] X. Jin, L. E. Li, L. Vanbever, J. Rexford. "Softcell: Scalable and Flexible Cellular Core Network Architecture," in *Proceedings of the Ninth ACM Conference on Emerging Networking Experiments and Technologies*, ACM, California, CA, USA, December 2013, pp. 163-174.
- [7] TNOVA EU project, <http://www.t-nova.eu/>.
- [8] UNIFY EU project, <http://www.fp7-unify.eu/>.
- [9] A. L. Valdivieso Caraguay, L. I. Barona López, L. J. García Villalba. "Evolution and Challenges of Software Defined Networking," in *Proceedings of the Workshop on Software Defined Networks for Future Networks and Services*, Trento, Italy, November 2013, pp. 47-55.
- [10] A. L. Valdivieso Caraguay, A. Benito Peral, L. I. Barona López, L. J. García Villalba. "SDN: Evolution and Opportunities in the Development IoT Applications," *International Journal of Distributed Sensor Networks*, IJDSN, May 2014, pp. 1-10.
- [11] N. McKeown, T. Anderson, H. Balakrishnan, et al. "OpenFlow: Enabling Innovation in Campus Networks," in *ACM SIGCOMM Computer Communication Review*, vol. 38, no.2, April 2008, pp. 69-74.
- [12] ETSI Industry Specification Group (ISG). "Network Function Virtualization (NFV) White Paper," in *SDN and OpenFlow World Congress*, Frankfurt, Germany, September 2013, pp. 1-16.
- [13] H. Hawilo, A. Shami, M. Mirahmadi, R. Asal. "NFV: State of the Art, Challenges, and Implementation in Next Generation Mobile Networks (vEPC)," in *IEEE Network*, vol. 28, no. 6, November 2014, pp. 18-26.
- [14] S. Costanzo, L. Galluccio, G. Morabito, S. Palazzo. "Software Defined Wireless Networks: Unbridling SDNs," in *Proceedings of the European Workshop on Software Defined Networking*, EWSDN, Darmstadt, Germany, October 2012, pp. 1-6.
- [15] C. Chaudet, Y. Haddad. "Wireless Software Defined Networks: Challenges and Opportunities," in *Proceedings of the 2013 IEEE International Conference on Microwaves, Communications, Antennas and Electronics Systems*, COMCAS, Tel Aviv, Israel, October 2013, pp. 1-5.
- [16] K. K. Yap, R. Sherwood, M. Kobayashi, T. Y. Huang, M. Chan, N. Handigol, G. Parulkar. "Blueprint for Introducing Innovation into Wireless Mobile Networks," in *Proceedings of the Second SIGCOMM Workshop on Virtualized Infrastructure Systems and Architectures*, ACM, New Delhi, India, September 2010, pp. 25-32.
- [17] A. Detti, C. Pisa, S. Salsano, N. Blefari-Melazzi. "Wireless Mesh Software Defined Networks (wmSDN)," in *Proceedings of the 9th IEEE International Conference on Wireless and Mobile Computing, Networking and Communications*, WiMob, Lyon, France, October 2013, pp. 89-95.
- [18] P. Dely, A. Kessler, N. Bayer. "Openflow for Wireless Mesh Networks," in *Proceedings of the 20th IEEE International Conference on Computer Communications and Networks*, ICCCN, Maui, Hawaii, July 2011, pp. 1-6.
- [19] A. Basta, W. Kellerer, M. Hoffmann, K. Hoffmann, E.-D. Schmidt. "A Virtual SDN-Enabled LTE EPC Architecture: A Case Study for S-/P-Gateways Functions," in *Proceedings of the Workshop on Software Defined Networks for Future Networks and Services*, Trento, Italy, November 2013, pp. 8-14.
- [20] S. B. H. Said, M. R. Sama, K. Guillovard, L. Suci, G. Simon, X. Lagrange, J. M. Bonnin. "New Control Plane in 3GPP LTE/EPC Architecture for On-Demand Connectivity Service," in *Proceedings of the Second IEEE International Conference on Cloud Networking*, CloudNet, San Francisco, SF, USA, November 2013, pp. 205-209.
- [21] K. Pentikousis, Y. Wang, W. Hu. "Mobileflow: Toward Software-defined Mobile Networks," in *IEEE Communications Magazine*, vol. 51, no. 7, July 2013, pp. 44-53.
- [22] R. Riggio, T. Rasheed, F. Granelli. "EmPOWER: A Testbed for Network Function Virtualization Research and Experimentation," in *Proceedings of the Workshop on Software Defined Networks for Future Networks and Services*, Trento, Italy, November 2013, pp. 138-142.

- [23] R. Sherwood, G. Gibb, K. K. Yap, G. Appenzeller, M. Casado, N. McKeown, G. M. Parulkar. "Can the Production Network be the testbed?," in *Proceedings of the 9th USENIX Symposium on Operating Systems Design and Implementation*, OSDI, Vancouver, BC, Canada, vol. 10, October 2010, pp. 365-378.
- [24] A. Gudipati, D. Perry, L. E. Li, S. Katti. "SoftRAN: Software Defined Radio Access Network," in *Proceedings of the Second ACM SIGCOMM Workshop on Hot Topics in Software Defined Networking*, ACM, Hong Kong, China, August 2013, pp. 25-30.
- [25] L. E. Li, Z. M. Mao, J. Rexford. "Toward Software-Defined Cellular Networks," in *Proceedings of the European Workshop on Software Defined Networking*, EWSDN, Darmstadt, Germany, October 2012, pp. 7-12.
- [26] L. E. Li, Z. M. Mao, J. Rexford. "CellSDN: Software-Defined Cellular Networks," in *Technical Report*, Princeton University, 2012.
- [27] ICT-CROWD EU project, <http://www.ict-crowd.eu/>.
- [28] CITYFLOW project, <http://www.onesource.pt/cityflow/site/>.
- [29] OFELIA EU project, <http://www.fp7-ofelia.eu/>.
- [30] EUKOREA EU project, <http://eukorea-fire.eu/pilots/>.
- [31] F. Schneider, T. Egawa, S. Schaller, S. I. Hayano, M. Schöller, F. Zdarsky. "Standardizations of SDN and Its Practical Implementation," in *NEC Technical Journal*, vol. 8, no. 2, April 2014, pp. 16-20.

Impact of Node Clustering on Power Consumption in WSN

A Comparative Study

Ala'a H. Makableh, Ghassan Samara
Department of Computer Science
Zarqa University
Zarqa, Jordan

Abstract— Wireless Sensor Networks (WSN) is an important application uses the power of wireless communication to querying the real world. Sensor nodes are battery-driven devices and it have to work as long as possible to gather data. This paper provides a comparative study about using clustering on WSNs and how it helps in saving energy in these sensor nodes. The aim of this study is suggesting a topology to distribute nodes in WSN in a way that enhance overall battery life.

Keywords — Wireless Sensor Networks, Clustering, Power Consumption.

I. INTRODUCTION

Wireless Sensor Networks (WSN) is a network of sensor nodes, each node works as station to collect specific data from its environment, organizing these collected data and sent them to central computer as electrical signals wirelessly to be manipulated. In WSN, sensors send these signals either periodically or based on events depending on the goal of censoring.

The importance of WSN comes from its critical using in many applications as in [1] and [2] some of these applications can be divided to fields as follows:

- Military Applications for targeting or detecting the Nuclear, biological and chemical attack.
- Environmental Application like detecting the forests fire, flooding and air or water pollution.
- Health Applications like tracing and monitoring patient in the hospital and drug administration.
- Home applications in which the home machines will be controlled remotely by end user and interacts automatically.

The main component of wireless sensor node consist of:

- Sensing Unit, which is a hardware that responsible about measuring the physical parameters.
- Processing Unit, which is a hardware that process the data, collected by sensor unit and controls the functionality of other components in the sensor node.
- Transceiver Unit, which is a device work as a transmitter and a receiver at the same time.
- Power Unit, which is typically the battery.

Some optional component may be added such as Location Finding System, Power Generator, Mobilizer. A simple structure of sensor node shown in Figure 1.

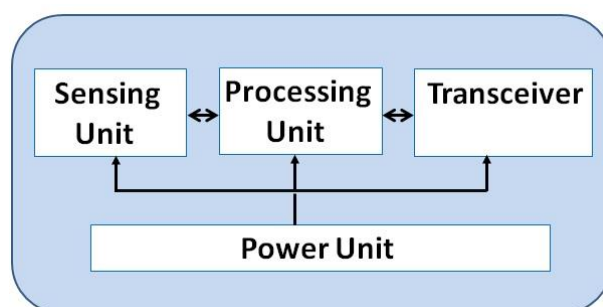


Fig. 1. The main component of wireless sensor node

Since the nodes in WSN are battery-driven and battery is a limited power source, it is important to minimize the energy that the node consume to maximize the overall efficiency of WSN.

The constraints on WSN which come from the limited energy, storage size and processing power leads to take these constraints in account and adapt the wireless network protocols or suggests a new protocols specifically designed for wireless sensor networks to achieve these requirements.

In this paper the power consumption and battery limitation for nodes in WSN is studied, furthermore, this paper will show that clustering can enhance the overall power consumption and provides a topology that exploit the node clustering to enhance power consumption.

The work is organized as follows. In section II literature review is presented. In section III clustering techniques are discussed and compared. A new model for WSN clustering suggested in section IV. In section V the conclusion presented.

II. LITERATURE REVIEW

The previous researches in WSN covered many ideas to solve power consumption and suggested different models. In [3] the power consumption studied from the point of using multi-hop implementation in WSN to reduce the power consumption for the sensor nodes. In contrary, an opposite result suggested in [4] where a single-hop implementation is used which is simpler in routing protocol and needs a lower communication overhead which is more efficient. Another idea proposed in [5] to serve the applications in environmental science and agriculture by distributing large number of sensor nodes in wireless network which combines a very large radio range with low cost, and transmits only in a single hop mode to get low power consumption. The improvement of hardware and implementing a hardware-based forward error correction scheme is suggested in [6] which gives a better overall energy consumption for the node.

As mentioned in [7] routing protocols classified into seven main categories: Heterogeneity-based Protocols, Location-based Protocols, Data centric Protocols, Hierarchical Protocols, Mobility-based Protocols, Multipath-based Protocols and QoS-based Protocols. Hierarchical Protocols organizes the nodes in WSN as a small group named cluster.

The cluster has two-level hierarchy of nodes: the first higher is the cluster head nodes collecting data from nodes in its cluster and transmit these data, the second lower is the node members of the cluster which are responsible for collecting the real world data [8]. This saves communication and processing work and also saves energy [9]. The importance of energy efficiency not only in static sensor nodes but also in mobile sensor networks too. One of the modern researches studied the mobile sensor networks and proposed a novel approach to develop an energy efficient routing in [10].

III. CLUSTERING TECHNIQUES:

A part from clustering architecture of a WSN is treated to minimize the energy consumption by transmitting less data, and this also improves the scalability of the network and the communication bandwidth within the cluster [10]. Some of the energy efficient routing protocols based on clustering are LEACH, HEED, PEGASIS, Hierarchical-PEGASIS, TEEN, APTEEN:

Low Energy Adaptive Clustering Hierarchy (LEACH) [11], this protocol uses the clustering to distribute the energy consumption by dividing network nodes into groups based on data collection. The cluster heads that collect data from the nodes coming under its cluster are randomly selected. This random selection helps to prevent energy draining for the same sensor node because head would be changed. The main problem

of this protocol is the random selection of cluster head. In the worst case the choose of cluster head nodes may be not distributed fairly, and this effects the data gathering. See figure 2 which shows the LEACH where the cluster heads collect data before transmitting it to the base station (BS) [12].

Hybrid Energy Efficient Distributed clustering approach (HEED) [13] was developed to avoid the problem of random selection of cluster heads selects the cluster heads based on both remained energy level and communication cost.

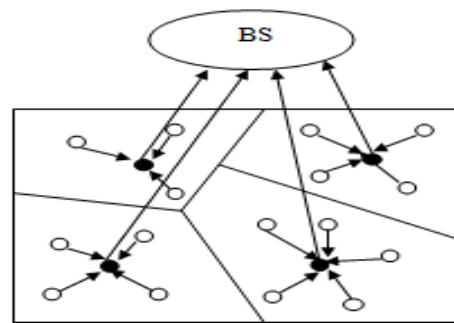


Fig. 2. LEACH Architecture

These protocols are proved energy efficient to WSN with static sensor nodes only and they are untested while the sensor nodes exhibit mobility [10].

Power Efficient Gathering in Sensor Information Systems (PEGASIS), where all the sensor nodes in the network will be arranged to form a chain with a leader node which is responsible for transmitting data to the base station [14]. The data moves from one node to the next until reaching the end of the chain where the leader node lies, the leader node transmits one message to the base station as shown in Figure 3 [15]. The main problem of this protocol is the long delay.

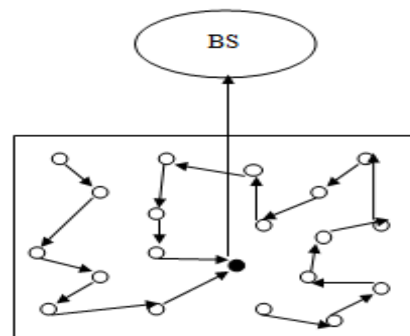


Fig. 3. PEGASIS Architecture

Hierarchical-PEGASIS is an enhancement on PEGASIS, to reduce transmission delay [16].

Threshold Sensitive Energy Efficient Sensor Network Protocol (TEEN) is a hierarchical clustering based protocol in which nodes react with the changes in the environment. After

cluster formation, the cluster head broadcasts two thresholds to the sensor nodes namely hard threshold and soft threshold. Hard threshold permits the sensor nodes to send data only when the attribute sensed by them is in the range of interest. The soft threshold will reduce the data transmission if there is no or little change in the value of sensed attribute. In order to control the data transmissions, both thresholds can adjust [17].

The Adaptive Threshold sensitive Energy Efficient sensor Network protocol (APTEEN) is an extension to TEEN. Clustering in APTEEN and TEEN protocols is done in two levels of cluster heads: the first level collects data from set of simple nodes and the second level cluster heads gets data from cluster heads in the first level and its simple nodes to transmit it to base station as shown in Figure 4.

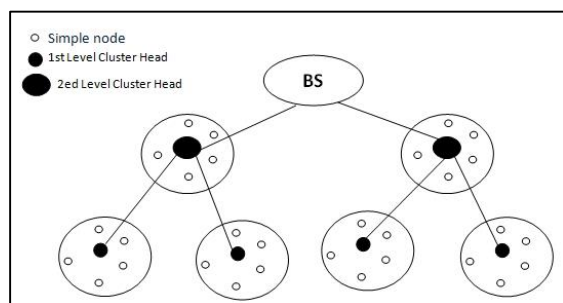


Fig. 4. Hierarchical Clustering in TEEN

A summary to these protocols can be as follows [13]:

LEACH, TEEN, APTEEN and PEGASIS have almost similar features and fixed infrastructure. LEACH, TEEN, APTEEN are cluster based routing protocols but PEGASIS is a chain-based routing protocol. The performance of APTEEN is intermediate between TEEN and LEACH in the field of power consumption of the network. PEGASIS avoids the overhead of cluster formation of LEACH, but it needs dynamic topology adjustment. PEGASIS adds excessive delay for distant nodes on the chain.

IV. PROPOSED SYSTEM

Based on nature we propose a perfect hexagon clustering system to the WSN. This system takes the advantages of hexagon shape like what the bee do in beehive to get maximum amount of space with a minimal amount of material.

When the area of the cluster of WSN is divided into a grid of hexagon the resulted grid will have no gaps and it covers the maximum area unlike the other shapes such as the circle or octagon create gaps and the triangle or square makes the area smaller as shown in figure 5.

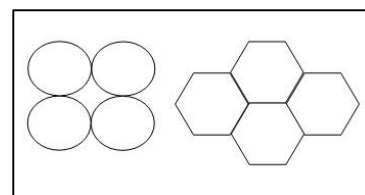


Fig. 5. No gaps in hexagon grid

In the cluster, each hexagon will be called cluster cell. The cluster cell will have sensor node in the central, this node responsible about sensing within its cell area. By applying this to all cluster cells the whole cluster will cover the maximum area and the distances between these nodes will be minimized to get efficient communication with low power consumption within the nodes of the same cluster. Figure 6 shows the architecture of such system.

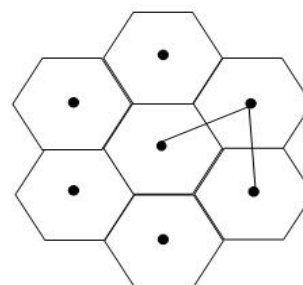


Fig. 6. a perfect hexagon cluster with centric sensor nodes

V. CONCLUSION

In WSN it is still important to maximize the network nodes efficiency with minimization to the power consumption. This paper studied the clustering protocols and suggested a new clustering model. This model enhances the cluster efficiency by maximizing the sensing area with better communication distances to minimize the power consumed in communication between nodes.

REFERENCES

- [1] John A. Stankovic, Anthony D. Wood, Tian He, "Realistic Applications for Wireless Sensor Networks", Theoretical Aspects of Distributed Computing in Sensor Networks, Springer Verlag, 2010.
- [2] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless Sensor Networks: A Survey", Elsevier Computer Networks, March 2002.
- [3] J. M. Rabaey, J. Ammer, T. Karalar, S. Li, B. Otis, M. Sheets, T. Tuan, "Pico Radios for wireless sensor networks—the next challenge in ultra-low power design, International Solid-State Circuits Conference (ISSCC), 2002.

- [4] Zhong, L.C.; Rabaey, J.M.; Wolisz, “ Does proper coding make single-hop wireless sensor networks reality: the power consumption perspective”, *Wireless Communications and Networking Conference (WCNC) IEEE*, 2005.
- [5] C. Huebner, R. Cardell-Oliver, S. Hanelt, T. Wagenknecht, A. Monsalve, “Long range wireless sensor networks with transmit-only nodes and software defined receivers”, *Canadian Society of Safety Engineering (CSSE) Technical Report*, July 2010.
- [6] Andreas Brokalakis, Ioannis Papaefstathiou, “Using hardware-based forward error correction to reduce the overall energy consumption of WSNs”, *Wireless Communications and Networking Conference (WCNC) IEEE*, 2012.
- [7] Parul Tyagi, Surbhi Jain, “Comparative Study of Routing Protocols in Wireless Sensor Network”, *International Journal of Advanced Research in Computer Science and Software Engineering*, 2012.
- [8] Vivek katiyar, Narottam chand, Surender soni, “A Survey on Clustering Algorithms for Heterogeneous Wireless Sensor Networks”, *International Journal of Advanced Networking and Applications*, 2011.
- [9] Ameer Ahmed Abbasi, Mohamed Younis, “A survey on clustering algorithms for wireless sensor networks”, *IEEE Communication Magazine*, 2007.
- [10] B.Baranidharan, B.Shanthi, “A Survey on Energy Efficient Protocols for Wireless Sensor Networks”, *International Journal of Computer Applications (0975 – 8887)*, 2010.
- [11] W. B. Heinzelman, A. P. Chandrakasan, H. Balakrishnan, “Application specific protocol architecture for wireless micro sensor networks”, *IEEE Transactions on Wireless Networking* 2002.
- [12] Anjali Bharti, Kanika Sharma, “Comparative Study of Clustering based Routing Protocols for Wireless Sensor Network”, *International Journal of Computer Applications (0975 – 8887)*, 2013.
- [13] O. Younis, S. Fahmy, “HEED: A Hybrid Energy-Efficient Distributed clustering approach for Ad Hoc sensor networks”, *IEEE Transactions on Mobile Computing* 3, 2004.
- [14] Lindsey, S. Raghavendra, C. S., “PEGASIS: Power Efficient gathering in sensor information systems”, *Proceedings of IEEE Aerospace Conference*, 2002.
- [15] Jisoo Shin, Changjin Suh, “CREEC: Chain Routing with Even Energy Consumption”, *IEEE Journal of Communications and Networks*, 2011.
- [16] S. Lindsey, C. S. Raghavendra, K. Sivalingam, “Data gathering in sensor networks using the energy*delay metric”, *IEEE International Parallel & Distributed Processing Symposium (IPDPS) Workshop on Issues in Wireless Networks and Mobile Computing*, 2001.
- [17] Manjeswar, A. Agrawal, D. P., “TEEN: A protocol for enhanced efficiency in wireless sensor networks”, *International Workshop on Parallel and Distributed Computing Issues in Wireless Networks and Mobile computing*, 2001.

Frequency Assignment for Cellular Mobile Systems Using a Memetic Algorithm

Tounsi Abdelkader¹, Babes Malika²

^{#1} University Badji Mokhtar, Department of Computer Science, Laboratory of networks and system
BP 12 23000 Annaba, Algeria
tounsi.abdelkader@hotmail.com

^{#2} University Badji Mokhtar, Department of Computer Science
BP 12 23000 Annaba, Algeria,
malikababes@yahoo.fr

Abstract —This paper presents a novel algorithm for finding a solution to the frequency assignment problem (FAP). The algorithm proposed here is a metaheuristic approach called memetic algorithm, which uses genetic algorithm based on new crossover operator combined with a modified tabu search. The performance of our algorithm is evaluated using for the results eight well-known benchmark problems, Computational results allow us to confirm the effectiveness of the proposed algorithm.

Keywords: frequency assignment problem; genetic algorithm; tabu search; memetic algorithm, metaheuristic.

I. INTRODUCTION

Due to the insufficiency of available bandwidth resources and the increasing demand for cellular communication services, the Frequency assignment problem becomes increasingly important.

Frequency assignment is generalization of the graph coloring problem, which belongs to the class of NP-complete problems [4]. For this category of problem, there is no known algorithm that can generate a guaranteed optimal solution in an execution time that may be expressed as a finite polynomial of the problem dimension. As this reason many heuristic methods have been proposed to deal with FAP. This large production has been analyzed and organized in several surveys [1, 2, 3]. We briefly present about each of used approaches:

- genetic algorithm based approaches.
- Constructive algorithms and local searches.
- Tabu Search.
- hybrid algorithm.
- approach based on ant colony paradigms.
- Neural Network Algorithm
- Simulated annealing
- Multi agent Systems

From the experiments made by most researchers on a variety of standard benchmark problems, tabu search(TS) has been one of the most effective heuristic algorithms for the FAP especially when solution time is less important than solution quality [5]. The differences in these approaches lie in the representation of the move, in the definition of neighborhood of a move, and in the way of defining a tabu move [1]. Moreover genetic

algorithm (GA) provides for a wide exploration of the space to search. As these reasons the hybridization between the two last approaches is meant to accelerate the discovery of good solutions.

Memetic algorithms (MAs), which are similar to genetic algorithms, are good algorithms for combinatorial optimization problems [14]. Normally, a genetic algorithm combined with local search methods is called a memetic algorithm (MA) [14]. MAs have received various names throughout the literature and scientists not always agree what is and what is not an MA due to the large variety of implementations available. Some of the alternative names used for this search framework are hybrid GAs, Baldwinian EAs, Lamarckian EAs, genetic local search algorithms, and other names are not unheard of. Moscato [13] coined the name Memetic Algorithm to cover a wide range of techniques where evolutionary-based search is augmented by the addition of one or more phases of local search.

In this paper, we propose a memetic approach for solving the frequency assignment problem. We consider a general cellular radio network subjected to all three kinds of constraints [8] :

1. the cochannel constraint (CCC): the same frequency cannot be assigned to certain pairs of radio cells simultaneously;
2. the adjacent channel constraint(ACC): frequencies adjacent in the frequency domain cannot be assigned to adjacent radio cells simultaneously;
3. the co-site constraint (CSC): any pair of frequencies assigned to a radio cell must have certain distance in the frequency domain.

The rest of the paper is organized as follows. In Section 2, we formulate the frequency assignment problem. In Section 3, we present the proposed algorithm. In Section 4, experimental results are presented. Future work and perspectives are discussed in the last section.

II. PROBLEM DESCRIPTION

The problem of frequency assignment is to provide wireless communication frequencies from limited spectral resources while keeping to a minimum interference suffered by those wishing to communicate in a given radio communication network. According to [2] the here considered version of FAP is classified as a MS-FAP (minimum Span frequency assignment problem). Radio channels are represented by the positive integers. Let $M = \{1, 2, \dots, m\}$ be a set of available channels where m is the Number of available channels in the mobile network. The basic model of the channel assignment problem can be represented as follows:

1. N : The number of cells in the mobile network.
2. d_i : represents the number of frequencies that must be assigned to cell i ($1 \leq i \leq N$)
3. C : Compatibility matrix, $C=(c_{ij})_{N \times N}$ represents the Minimal channel separation between channels in cells i and j , $1 \leq i, j \leq N$.
4. $Call_{ik}$: Cell i with call k where $1 \leq i \leq N$, $1 \leq k \leq d_i$
5. f_{ik} : A radio channel is assigned to $Call_{ik}$, where f_{ik} belongs to the set of radio channel F .
6. *Frequency separation constraint* :
 $|f_{ik} - f_{jm}| \geq c_{ij}$, for all i, j, k, m ($i \neq j, k \neq m$), c_{ij} is defined in Compatibility Matrix, C . If ($i=j$), it become co-site constraint.
7. *TotalAssignCh*: the total number of required channels The total of radio channel to be assigned in the system can be shown as:

$$\text{TotalAssignCh} = \sum_{i=1}^n d(i) \quad (1)$$

Therefore, the objective is to find an assignment that minimizes the total number of violations in an assignment, Subject to:

$$\min \sum_{i=1}^n \sum_{k=1}^m \sum_{j=1}^n \sum_{l=1}^m f_{ik} \text{viol}(i,j,k,l) f_{jl} \quad (2)$$

Where: $\text{viol}(i,j,k,l) = \begin{cases} 0 & \text{if } |k-l| \geq C_{ij} \\ 1 & \text{otherwise} \end{cases}$
 and

$$f_{ik} = \begin{cases} 0 & \text{if channel } k \text{ is not assigned to cell } i \\ 1 & \text{otherwise} \end{cases}$$

for $1 \leq k, l \leq m$ and $1 \leq i, j \leq N$.

III. MEMETIC ALGORITHM

In this paper, a Memetic Algorithm (MA) is proposed for FAP. The mechanisms and operators of the MA are given below.

A. Representation and selection mechanisms

Encoding the chromosome in FAP is relatively simple, We represent a solution by a binary matrix $N \times m$. N is the number of cells and m is the number of channels. If a gene $f_{ik} = 1$, then the k^{th} channel is assigned to the i^{th} cell. Chromosomes are randomly selected from the population for crossover and mutation.

B. Population initialization

In a traditional GA, each chromosome in the initial population is generated randomly. However, many experimental results have shown that the GA does not always result in good solutions due to the random method.

the co-site constraint (CSC) may cause more interference in the real-world situation[6]. So firstly to make sure the absence of interference between two channels assigned to the same cell, we have generate the initial population with new method called (PI) inspired from[7]. Figure 1 shows the difference between the two methods: -Cell: i

- j : is random number between 1 and $m - ((d_i - 1) \times C_{ii} + 1)$

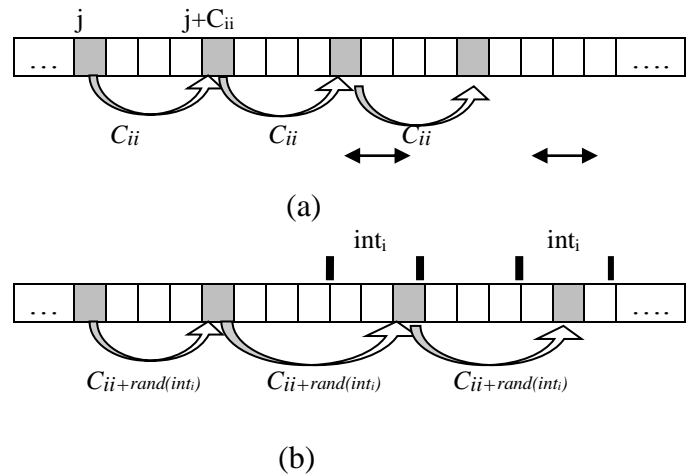


Fig.1: Frequencies assigned to cell 'i'
 (a) Cheng method [7] (b) our method(PI)

The int_i is defined by

$$\text{Int}_i = \frac{m - ((d_i - 1) \times C_{ii} + 1)}{d_i} \quad (3)$$

Where:

C_{ii} represents the Minimal channel separation between channels in cell i ;

$\text{Int}_i + C_{ii}$ is the maximum channel separation between channel in cell i ;

Int_i represents the interval of possible assignment to cell "i" without breaking the CSC constraint of all assignment in the Cell i .

The cells with a high channel demand have more difficulty obtaining an assignment that satisfies the CSC constraints. So we have generated p_i to deal with this problem. Where p_i is defined as:

$$p_i = \frac{(d_i - 1) \times C_{ii} + 1}{m} \quad (4)$$

Cells with a high value of p_i means that these cells must use the minimum separation distance (C_{ii}) between channels. For the others with low value of p_i means that these cells use the separation distance $C_{ii} + \text{random}(0, \text{int}_i)$. Figure 2 is the pseudo code of PI. Where sol_i is the assignment of cell i .

C. Proposed crossover operator

Chromosomes in the population have good genes resulting from the population initialization, It is clear that performance can be improved by transferring these good genes to the next generation.

An overview of the crossover operator is shown by figure 3. Two parents such as parent1 and parent2 are considered for the crossover process. We randomly generate a binary number mask then, we calculate P_i for each cell in the two parents where P_i is defined in (4). The value of P_i can exclusively reduce the searching space and consequently the convergence time is shortened. P_i has two roles; the first one is keeping CSC constraints for some cells verified, the other one is for the cells that do not have a high probability for improvement - optimum solution obtained - will be leaved uncrossed.

Population initialization

for each cell i

$a = \text{random}(0, 1)$

if ($a > p_i$)

for each demand d_j of cell i

$\text{sol}_i(d_j) = [\text{random}(0; \text{Int}_i) + \text{sol}_i(d_{j-1}) + C_{ii}] \text{mod } m + 1$

end

else

$s = \text{random}(m - (r_i - 1) \times C_{ii})$

$\text{sol}_i(1, \dots, d_i) = [s, s + C_{ii}, s + 2 \times C_{ii}, \dots, s + (d_i - 1) \times C_{ii}]$

endif

end for

Fig. 2: population initialization

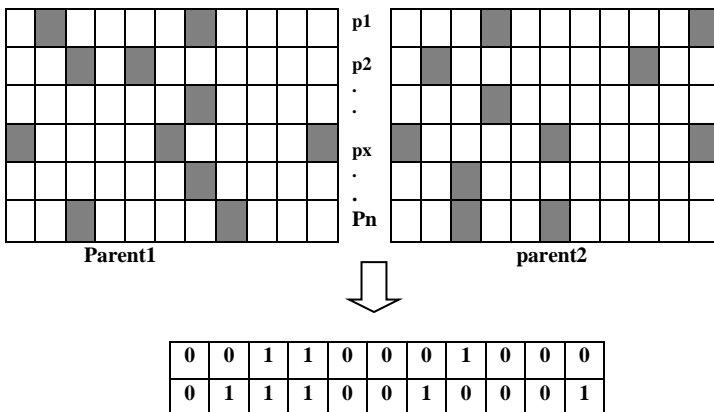
$C = \text{Random}(0, 1)$

If ($C < p_i$)

The row pairs leave uncrossed

Else

The row pairs will be crossed



1	0	0	0	1	1	1	0	1	0	1
0	0	0	0	0	0	0	0	0	0	0
1	0	1	1	0	1	1	0	1	0	0
1	1	0	0	1	1	0	1	0	0	0

($C < P_x$)

Binary mask

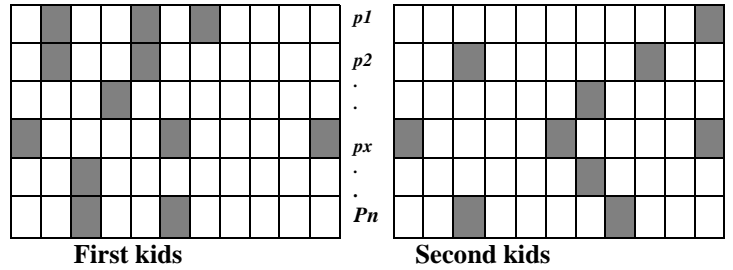


Fig. 3: Proposed crossover operator

D. Mutation operator

Our mutation operator called "mutation adapter" controls the total number of "1" in row i . we make sure that not to exceed d_i of the demand vector in any case. We do this by adding or removing a "1" randomly from the specified row. The mechanism of mutation operator is shown in figure 4.

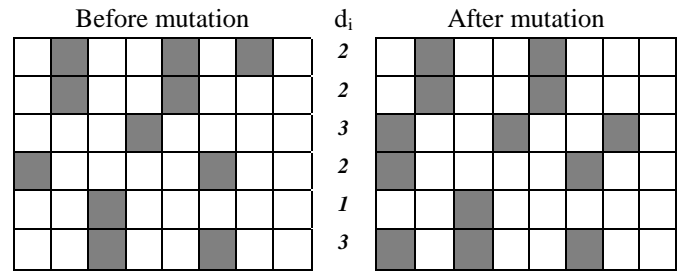


Fig. 4: mutation adapter

E. MA description

Figure 5 illustrates the MA framework followed in the implementation of the proposed algorithm. The first step of the algorithm is to create an initial population with PI and evaluate it. After that, for each generation, a pair of "parent" solutions is selected for the above methods of crossover and mutation. Two solutions are created which typically shares many of the characteristics of its "parents". After that the Tabu search and the module of restricted frequency are performed on these solutions. The mechanisms and operators of these components are given in section F and G.

Algorithm MA

	Fx	fy				
Interference ↑ cell=i ↓ cell=j
	.	1

	.	1	.	0

embrace more efficient and systematic forms of direction such as memorizing and learning [15].the hybridization between GA and TS is more successful in our case.

1) *Solution representation*

The representation of a frequency assignment S is obtained by using a matrix $N \times m$. where if a gene $f_{ik} = 1$, then the k^{th} channel is assigned to the i^{th} cell.

2) *Neighborhood*

The space to search or set of moves is defined in figure 6. Our basic move is then moving a 1 to a 0 entry of the same row. Thus changing the assigned channel of one cell, maintaining the number of assigned channels unchanged and keeping the result of crossover operator (co-site interference).

3) *Tabu list*

Some neighbors composed of previously encountered solutions will not be considered for the next k iterations (k , called tabu length). These neighbors are consisted in tabu list, which is one of the main mechanisms of tabu search. Tabu list is also the most important feature distinguished with other search algorithms. Thus, TS can be described as a form of neighborhood search with a set of critical and complementary components.

As shown in Figure 6, the element (j, fx) prevents the algorithm from re-visiting previously seen solutions. The element (j, fy) prevents the other individuals from re-exploring the same search area.

.
.	1
.
.
.	0	.	1

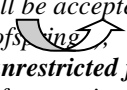
Begin
 Initialize population P with PI;
 Evaluate initial population,
 Repeat
 -Generate-Binary-Masks
 -Select two parents randomly
 -use the **proposed crossover**. Then get two offspring
 - **mutation adapter**(offspring1, offspring2)
 - If fitness (offspring2) better than fitness (parent)
 then Offspring will be accepted.
 - **tabu Search** (ofspring) 
 - calculate the **unrestricted frequencies** for each cell which have interference in offspring1
 - replace the restricted frequencies with unrestricted frequencies
 -add offspring1 to population
 Until Termination criteria are satisfied;
 End;

Fig. 5: MA framework

F. *Tabu search*

After several generations, chromosomes in the population are similar enough to each other such that only local optimization may be possible. Tabu search is one of the mechanisms to avoid minimum local. Opposite to randomizing approaches such as Simulated Annealing (SA) where randomness is widely used, TS is based on the principle that intelligent search must


 **tabulist = (j,fx); (j,fy).**

Fig. 6: Tabu search moves

G. *Restricted frequencies*

In order to improve the quality of the solution obtained, we apply the following steps to locate the free frequencies (if any exist) and resolve the maximum of conflict between cells. This

module is applied to solution obtained by tabu search that do not satisfy all the constraints of the problem.

1. determine the couples of cells which have interference between them ;
2. from the list of couples “S” ,we create list of candidate L .where L is defined by:
 $L = \{x, p_x < p_y \forall (x,y) \in S\}$ p_x, p_y defined in (4)
3. delete all frequencies assigned to L
4. Assigned={}
5. calculate” free frequencies” it means the authorization of assigning frequencies in cell x, $\forall x \in (L\text{-Assigned})$ (see figure 7) if (L-Assigned)={ } go to 7;
6. if (free frequencies) then
 $x \in (L\text{-Assigned})$ Assign the first free frequency to “x”,
 Assigned={x};
 else
 Assign a frequency randomly to x;
 Assigned={x};
 endif ;go to 5.
7. End

This approach shares some similarities in the calculation of the free frequency in Frequency Exhaustive Strategy [16].

H. Termination criteria

In traditional GA, either computation time or the number of generations or solution improved is selected as termination criterion. In this paper, all of them are considered. If a solution cannot be improved any more in consecutive generations, the algorithm terminates. However, it is very time-consuming for problems, MA terminates when any of the three conditions is true.

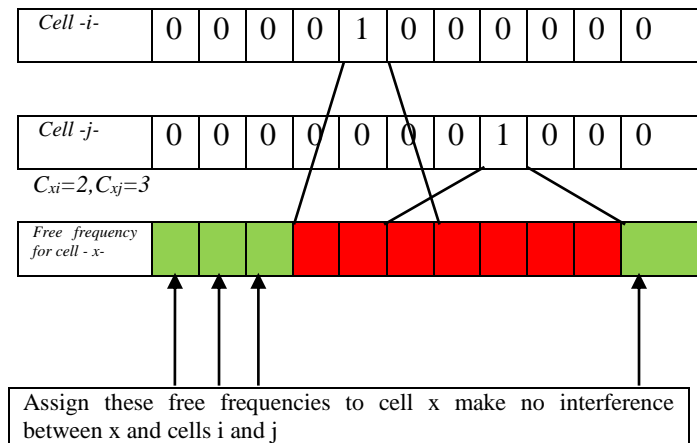


Fig. 7: Free frequencies

IV. EXPERIMENTAL RESULTS

In these experiments we used the Philadelphia problem instances [12]. These instances are widely known within problem FAP, and they are characterized by a number of hexagonal cells (habitually 21), that represent a cellular phone network of the Philadelphia city (see Figure 8).

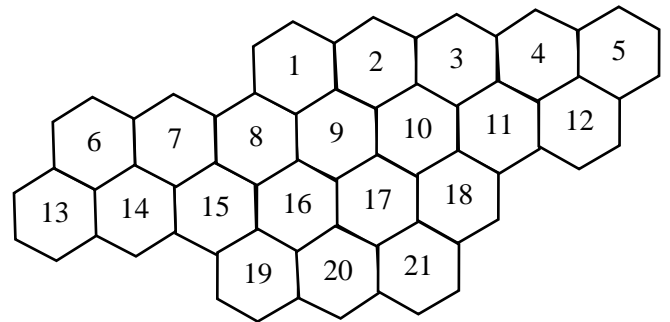


Fig. 8: Cellular geometry of test problems

The main differences among the instances correspond with the use of different interference matrices and demand vectors (see table 1). In these instances, the number of antennas (cells) varies from 4 to 25, and the number of frequencies finally used varies between 11 and 533. Figure 11 shows the different interference matrices (C) and the demand vectors (D) used.

Table I: Problem Specifications

Instance	# of Cell	Lower Bound (lb)	Compatibility matrix (C)	Demand vector (D)
P1	4	11	C_1	D_1
P2	25	73	C_2	D_2
P3	21	381	C_3	D_3
P4	21	533	C_4	D_3
P5	21	533	C_5	D_3
P6	21	221	C_3	D_4
P7	21	309	C_4	D_4
P8	21	309	C_5	D_4

To investigate the convergence frequency, Table II summarizes the simulation results, and shows the convergence to the optimum solution (the convergence frequency).

Table II: convergence rate

Inst	Approach and rate of convergence					
	Lower Bound	MA	[9] MGA	[10] NN	[11] DGuGA	[7] CSCP
P1	11	100%	100%	-	100%	100%
P2	73	100%	100%	62%	98%	100%
P3	381	100%	-	99%	-	100%
P4	533	100%	-	100%	-	100%
P5	533	100%	-	98%	-	100%
P6	221	100%	92%	97%	89%	100%
P7	309	100%	-	99%	-	100%
P8	309	100%	80%	52%	-	100%

Figure 9 shows the required generation to converge to the solution in two problems. In problem #8, our MA found the solution within 22 generation, in problem #2, our MA converge in the 36th generation.

Table 2 allows us to confirm the effectiveness of the proposed algorithm. The simulation results show that our algorithm achieved 100% convergence to solutions for all eight benchmarking problems. The results show that MA outperform the convergence results of the neural network (NN) [10], genetic algorithm (MGA) (DGuGA)[9,11]. The MA shows the same convergence results of [7]. Our algorithm identifies higher-quality solutions than other methods and it is easy to implement. For cellular systems with high number of cells, this algorithm can be efficiently applied to find the exact solution in an acceptable time of computation. Also since it is an implicitly parallel technique, it can be implemented very effectively on powerful parallel computers to solve exceptionally demanding large-scale problems.

Figure 10 shows a two-dimensional stem plot. It displays the frequency assigned to each cell for the problem #8, where the x-axis represent the cells, and the y-axis represent the frequency assigned. The full assignment of channels for a sample of the 8 benchmark problems considered in our simulations is presented in the Appendix. The results and analysis presented above indicate that our MA is certainly an effective Memetic algorithm for solving the frequency assignment problem.

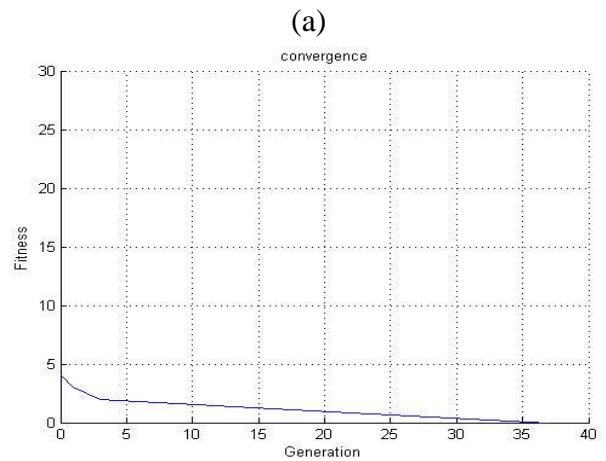
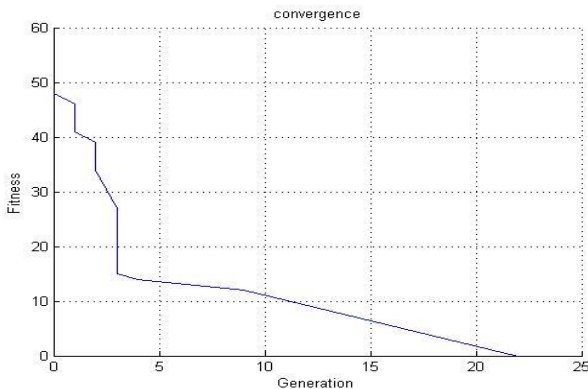


Fig. 9: (a)instance P8 convergence (b) instance P2 convergence

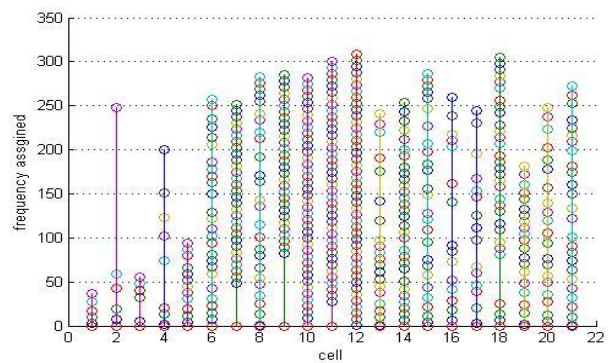


Fig. 10: two-dimensional stem plot

V. CONCLUSION

We have presented a memetic algorithm for solving the frequency assignment problem in cellular radio networks. Due the importance of CSC constraints, we have considered this in the population initialization. Problems specific information has been used into the mechanism of crossover and tabu search as to effectively bias the search process towards promising regions of the search space. After experiments with real data, optimization was achieved and the algorithm has been efficient and convergent. Our approach can find better or equivalent solutions compared with existing optimization methods. Our work presents a good perspective of managing the radio spectrum that can be intended towards a parallel implementation for implementing a pertinent strategy and effective management of resources in the future works.

Appendix:

This appendix shows the full assignments for benchmark instances No. 2, 5, and 6(see table 3-5)

Table III: Channel assignment for benchmark problem 2

Cells	Frequency
1	31,33,35,37,39,41,67,69,71,73
2	6,8,10,12,14,16,18,20,22,24,26
3	28,32,34,36,38,40,42,51,70
4	6,12,26,37,39
5	30,44,46,48,50,52,54,56,72
6	1,3,5,67
7	53,55,57,61,64
8	30,44,46,48,50,52,54
9	10,13,15,17
10	2,4,43,47,58,60,62,66
11	14,16,18,20,22,24,31,35
12	1,3,5,7,9,45,59,63,65
13	11,13,15,17,19,21,23,25,27,29
14	49,53,55,57,61,64,68
15	28,36,38,51,59,63,65
16	6,8,26,32,37,40
17	2,10,14,16
18	3,5,7,9,33
19	1,12,34,39,41
20	44,46,48,50,52,54,56
21	2,4,9,32,37,43
22	3,5,11,42
23	7,10,13,23,36
24	53,55,57,59,61,63,65
25	1,8,21,33,44

Table IV: Channel assignment for benchmark problem 5

Cells	Frequency
1	6,11,21,27,95
2	4,10,20,30,108
3	1,8,16,25,31
4	2,7,12,18,23,35,40,45
5	3,9,14,20,28,33,38,43,48,55,60,65
6	1,7,12,18,24,33,40,45,50,57,66,75,81,86,91,100,106,111,116,122,137,145,152,191,200
7	44,49,54,59,64,69,74,79,84,89,94,99,104,109,114,119,124,129,134,139,144,149,154,159,164,169,174,179,184,189
8	3,9,15,26,35,41,51,56,61,70,80,110,120,130,140,165,170,180,185,190,199,204,209,214,219
9	76,81,86,91,96,101,106,111,116,121,126,131,136,141,146,151,156,161,166,171,176,181,186,191,196,201,206,211,216,221
10	17,22,27,32,37,42,47,52,57,62,67,72,77,82,87,92,97,102,107,112,117,122,127,132,137,142,147,152,157,162,167,172,177,182,187,192,197,202,207,212
11	19,24,29,34,39,44,49,54,59,64,69,74,79,84,89,94,99,104,109,114,119,124,129,134,139,144,149,154,159,164,169,174,179,184,189,194,199,204,209,214
12	1,6,11,16,21,26,31,36,41,46,51,56,61,66,71,76,81,86,91,96,101,106,111,116,121,126,131,136,141,146,151,156,161,166,171,176,181,186,191,196,201,206,211,216,221
13	2,10,19,27,36,41,46,52,60,65,70,77,95,112,120,127,136,147,161,167
14	43,48,53,58,63,68,73,78,83,88,93,98,103,108,113,118,123,128,133,138,143,148,153,158,163,168,173,178,183,188
15	8,17,23,31,38,47,55,62,67,72,85,92,97,105,117,135,142,157,162,172,177,182,187,192,197
16	5,14,19,29,45,50,60,90,100,145,150,175,194,203,208
17	13,21,28,33,43,65,95,125,160,195,200,205,210,215,220
18	53,58,63,68,73,78,83,88,93,98,103,108,113,118,123,128,133,138,143,148,153,158,163,168,173,178,183,188,193,198
19	1,10,16,22,27,34,40,46,52,57,75,82,87,102,107,112,122,127,137,167
20	2,7,12,18,24,39,44,49,54,59,64,69,74,79,84,89,94,104,115,155
21	6,15,20,25,30,36,41,48,55,61,66,71,80,85,105,110,120,130,135,140,165,170,180,185,190

Cells	Frequency
1	10,17,31,51,61,68,89,271
2	6,19,46,59,76,96,103,118,131,139,150,164,180,187,194,201,208,220,227,234,241,248,255,262,269
3	11,26,54,62,109,122,129,167
4	7,16,42,68,142,236,272,326
5	34,83,137,170,194,220,248,286
6	1,11,22,29,45,60,69,76,88,97,161247,278,321,338
7	20,27,39,49,82,104,112,119,128,135,142,159,166,181,191,202,209,250
8	3,12,25,33,41,53,66,73,80,87,94,101,108,115,125,146,153,174,216,223,230,237,244,258,265,279,300,328,335,342,349,356,363,370,377,384,391,398,405,419,426,433,440,447,454,461,482,496,503,517,524,531
9	1,8,15,22,29,36,43,50,57,64,71,78,85,92,99,106,113,120,127,134,141,148,155,162,169,176,183,190,197,204,211,218,225,232,239,246,253,260,267,274,281,288,295,302,309,316,323,330,337,344,351,358,365,372,379,386,393,400,407,414,421,428,435,442,449,456,463,470,477,484,491,498,505,512,519,526,533
10	5,24,31,38,45,74,81,88,95,102,117,124,138,145,157,171,178,185,192,199,206,213,222,229,243,257,264,276
11	13,20,28,40,49,56,65,72,79,115,153,227,310
12	1,30,43,63,86,111,146,158,186,200,209,224,250,270,282
13	7,16,32,42,52,64,73,83,95,106,118,130,145,156,170,178,186,193,201,210,225,241,254,263,273,287,310,335,355,367,375
14	5,13,24,35,47,54,67,86,93,100,121,138,147,195,205
15	9,18,30,37,44,56,63,70,77,84,91,98,110,117,132,140,149,163,177,184,198,212,219,226,233,240,252,261,268,275,282,289,296,307,315,322,123,130,137,144,151,158,165,172,179,186,193,200,207,214,221,228,235,242,249,256,263,270,277,284,291,298,305,312,319,326,333,340,347,354,361,368,375,382,389,396,403,410,417,424,431,438,445,452,459,466,473,480,487,494,501,508,515
17	10,17,52,60,69,83,90,97,104,111,160,181,188,195,202,209,251,286,293,314,321,412,468,475,489,510,521,528
18	2,35,47,58,67,76,119,341
19	2,11,26,34,42,51,58,68,75,88
20	4,13,23,32,39,46,54,65,72,93,100,139,329
21	6,19,27,37,44,63,80,87,

Table V: Channel assignment for benchmark problem 6

C2	D2	C3	D3
2110101111011110000000000	10	511001111000011100000	8
1210101101011110000000000	11	151100111100001110000	25
1121111111111110000000000	9	115110011110000111000	8
0012001111111000000000111	5	011510001111000011000	8
1110200001111110000000000	9	001150000111000001000	8
0010021111000000000000000	4	100005110000111000000	15
1111012111111000000000000	5	110001511000111100100	18
1111011211111000000000010	7	111001151100011110110	52
1011011121110000000000011	4	111100115110001111111	77
111111112111111000001010	8	011110011511000111011	28
0011101111201111011111111	8	001110001151000011001	13
1111101111021100000000000	9	000110000115000001000	15
1111101101112111111100000	10	000001100000511000000	31
1110100001111211111100000	7	100001110000151100100	15
1100100001101121111111000	7	110001111000115110110	36
0000100001101112111100000	6	111000111100011511111	57
0000000000001111211000000	4	011100011110001151111	28
0000000000101111121100000	5	001110001111000115011	8
0000000000101111112111100	5	000000111000011110511	10
0000000000101111011211100	7	000000011100001111151	13
0000000000100010001121100	6	000000001110000111115	8
0000000001100010001112111	4		
0001000000100000001111211	5		
0001000111100000000011211	7		
0001000010100000000001112	5		

C4	D4	C5	C1	D1
711001111000011100000	5	721001221000011100000	5400	1
171100111100001110000	5	272100122100001110000	4501	1
117110011110000111000	5	127210012210000111000	0052	1
011710001111000011000	8	012720001221000011000	0125	3
001170000111000001000	12	001270000122000001000		
100007110000111000000	25	100007210000221000000		
110001711000111100100	30	210002721000122100100		
111001171100011110110	25	221001272100012210110		
111100117110001111111	30	122100127210001221111		
011110011711000111011	40	012210012721000122011		
001110001171000011001	40	001220001272000012001		
000110000117000001000	45	000120000127000001000		
000001100000711000000	20	000002100000721000000		
100001110000171100100	30	100002210000272100100		
110001111000117110110	25	110001221000127210210		
111000111100011711111	15	111000122100012721221		
011100011110001171111	15	011100012210001272122		
001110001111000117011	30	001110001221000127012		
000000111000011110711	20	000000111000012210721		
000000011100001111171	20	000000011100001221272		
000000001110000111117	25	000000001110000122127		

Fig. 11: Interference matrix (C) and demand vector (D) for the tested instances

REFERENCES

[1] Audhya, G. K., Sinha, K., Ghosh, S. C. and Sinha, B. P. (2011), A survey on the channel assignment problem in wireless networks. *Wireless Communications and Mobile Computing*, vol. 11 pp. 583–609

[2] K. I. Aardal, C.P.M. van Hoesel, A.M.C.A. Koster, C. Mannino, and A. Sassano, “Models and solution techniques for frequency assignment problems,” *Annals of Operations Research*, 2007, vol.153, pp.79–129.

[3] Katzela I., and Nagshineh M., Channel assignment schemes for cellular mobile telecommunication systems, *a comprehensive survey*, *IEEE Personal Communications*, 1996, pp.10–31.

[4] Hale WK Frequency assignment: theory and applications Proc IEEE , 1996 vol. 68(12) pp.1497–1514.

[5] R. Montemanni, J.N.J. Moon and D. H. Smith, “An improved tabu search algorithm for the fixed spectrum frequency assignment problem”, *IEEE Transactions on Vehicular Technology*, vol. 52(4), pp.891–901

[6] Yuanyuan Zhang; Ming Chen; , "A Metaheuristic approach for the Frequency Assignment Problem," *6th International Conference on Wireless Communications Networking and Mobile Computing (WiCOM)*, Sept. 2010, vol., no., pp.1-5, pp.23-25

[7] Cheng R-H., Yu C. W., Wu T-K, “A Novel Approach to the Fixed Channel Assignment Problem”. *Journal of Information Science and Engineering* 2005, 21, pp.39-58.

[8] A.Gamst, “Homogeneous distribution of frequencies in a regular hexagonal cell system,” *IEEE Trans. Veh. Technol*, Aug. 1982, vol. VT-31, pp. 132-144

[9] Ngo C. Y. and Li, V. O. K. “Fixed channel assignment in cellular radio networks using a modified genetic algorithm”, *IEEE Transactions on Vehicular Technology*, 1998, 47, pp.163–171.

[10] Kim J.-S., Park S. H, Dowd P. W., and Nasrabadi N. M. “Cellular radio channel assignment using a modified hopfield network”, *IEEE Transactions on Vehicular Technology*, 1997, 46, 4, pp.957–967.

[11] L.M. San Jose´-Revuelta “A new adaptive genetic algorithm for fixed channel assignment” *Information Sciences*, 2007 , vol. 177, pp. 2655-2678 .

[12] Eisenblätter, A., Koster, A. Web site in which the Philadelphia instances for the FAP problem are explained <http://fap.zib.de/problems/Philadelphia>, May 2007

[13] Moscato, P. (1989) “On Evolution, Search, ptimization, Genetic Algorithms and MartialArts: Towards Memetic Algorithms” *Caltech Concurrent Computation Program*, C3P Report 826.

[14] N. Krasnogor and J. Smith, “A Tutorial for Competent Memetic Algorithms: Model, Taxonomy and Design Issues,” *IEEE Transactions on Evolutionary Computation*, 2005, vol. 9, no. 5, pp. 474 - 488.

[15] Maninder Singh Kamboj, Jyotsna Sengupta, “Comparative Analysis of Simulated Annealing and Tabu Search Channel Allocation Algorithms”, *International Journal of Computer Theory and Engineering (IJCTE)*, 2009, Vol. 1, No. 5, pp. 588-591,

[16] Sivarajan, K.N., McEliece, R.J. and Ketchun, J.W. ‘Channel assignment in cellular radio’, *Proceedings of the 39th IEEE Vehicular Technology Conference*, May 1989, pp.846–850.

Application Layer Protocols to Protect Electronic Mail from Security Threats

Arwa Husien, Ghassan Samara
Department of computer science
Zarqa University
Zarqa, Jordan

Abstract—Electronic mail is the most widely used service from internet utilities, as it is experiencing phenomenal growth for personal uses and organizations. E-mail more valuable than phone for business communications, there are many threats for Electronic mail systems security such that phishing Electronic mail, spam, virus, spyware, and malware. Because of the nature of E-mail applications, Electronic mail security is a priority concern for many organizations and security practitioners face a unique set of management issues. Security levels, policies, privacy issues, confidentiality, message integrity. In this paper we discuss how to ensure the safety and security of corporate Electronic mail environment, detailing threats that should be prepared to avoid them, and tools that should be used to mitigate them.

Keywords— Electronic mail; Security; PGP; S/MIME; e-mail security; Security Protocols.

I. INTRODUCTION

Electronic mail is the most widely used and regarded service of network utilities, although it is very old service in the technology world, also Electronic mail still prevails as a significant business tool, E-mail systems have experienced phenomenal growth, from simple systems linking a few users on a single computer to vast international networks connecting correspondents on literally millions of different hosts.

However, there are some changes in using electronic mail systems over time, with more demands for mobile access-using wireless networks- and personal use the need of organizations today to keep their electronic mail systems secure due to the central role electronic mail plays in the modern enterprise.

Nowadays message contents are insecure, may be inspection by unauthorized people during its travelling in the network, there are Many corporate Electronic mail systems come with built-in security tools, but they are not nearly enough, According to experts at Trend Labs, the amount of Electronic mail considered bad jumped within the range of 88–90% of Electronic mail sent during the first three quarters of both 2010 and 2011.[4]

With the huge explosion of growing reliance on electronic mail for every essential and nonessential purpose, a demand for authentication and confidentiality services are grew rapidly. What users need is something more akin to standard mail (contents protected inside an envelope), they need to have confidence about the sender of the mail and its contents, as shown in figure(1) Electronic mail Encryption and Electronic mail Digital Signature are needed to achieve integrity and confidentiality in Electronic mail messaging.[2]

With more targeted threats across Network environment, how can the aspects of today's electronic mail services be protected?

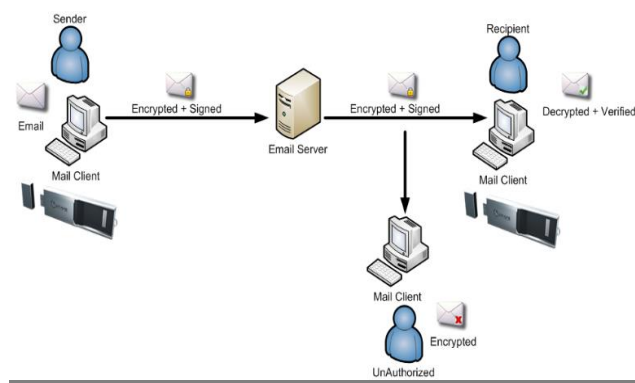


Fig.1.Electronic mail Security [8].

In network world we have a lot of application layer protocols for Electronic mail service such that Multipurpose Internet Mail Extension (MIME) which is an extension to the RFC 5322 protocol that is intended to solve problems and limitations of using Simple Mail Transfer Protocol (SMTP), which defined in RFC 821 which is traditional e-mail format standard, The most recent version of this format specification is RFC 5322, some of SMTP problems that it don't transmit all binary objects such that executable files, cannot transmit text data which includes national language characters, SMTP servers sometimes reject mail message over a certain size.

S/MIME is a security enhancement to the MIME Internet e-mail format standard based on technology from RSA which is Algorithm by Rivest, Shamir & Adleman for data security, also Pretty Good Privacy (PGP) is secure Electronic mail program, although both PGP and S/MIME are on an Internet Engineering Task Force (IETF) standards track, it appears that S/MIME used as industry standard for commercial and organizational use, while PGP used for personal e-mail .[2 , 3]

This paper discusses how users can ensuring the safety and security of corporate Electronic mail environment, detailing threats that should we prepare to avoid them, and tools may be used to mitigate them.

Electronic mail messages are the most cost-effective way to transmit information, as significant importance of electronic mail messaging and huge number of Electronic mail threads, such as spam, viruses and malwares, users need proper security measures to obtain the electronic mail security goals of confidentiality, message integrity, authentication, and non-repudiation from original.

Abbreviations and Acronyms

- PGP: Pretty Good Privacy.
- S/MIME: Security/ Multipurpose Internet Mail Extension.
- E-mail: Electronic mail.
- RSA: Rivest, Shamir & Adleman.

II. PREVIOUS STUDIES:

A. threats of email security

1) *viruses*: One of the most publicized and high risk of all the issues is viruses. Viruses are so dangerous ; they often deliver *highly fatal* load, destroying data, and dropping down entire mail systems.

Most of the viruses that were responsible for actual disasters during that time were either Internet worms or mass mailer viruses. To make matters worse, both of these virus types staying around longer than other types, even after anti-virus products have included protection against them. [11]

2) *SPAM*: Another major threat to email security today is SPAM (junk Email), often cited by organizations as being their number one concern, SPAM is considered a security threat because it can carry viruses, malicious code, and fraudulent solicitations for private information [11] . "junk email could cost a company with 500 employees nearly \$750,000 each year" [12] .

3) *Phishing*: Phishing (identity theft), is a newer threat to email security. Phishing is the process where identity thieves target customers of financial institutions, using common spamming techniques to generate huge numbers of emails with the intent of luring customers to spoofed web sites and Trapping them into giving personal information such as passwords. [11]

Phishing the most common methods of attack. Some of the threat and defenses are as follows masquerading: an attacker pretends to be someone else. In such situation, a criminal can set up a storefront and collect thousands or billions of credit card numbers from unsuspecting consumers. [10]

4) *The man in the middle*:The man in the middle attack and session hijacking attack occurs when an attacker inserts Itself between two parties and pretends to be one of the parties.

5) *Eavesdropping*: Eavesdropping happen when attacker listens to a private communication. The attacker views information as it is sent over the network. [10].

6) *Data diddling*: Data diddling attack happened when an attacker changes the data while it routing between communication parties.

7) *Dictionary attacks*: a dictionary attack happen when an attacker uses large set likely combinations to guess a

secret. aka, an attacker may choose one widely used password and try them all until the password is discovered.

8) *Denial of service attack*: denial of service attack occurs when an attacker floods the Email with hundred or even million of messages. Though the attacker does not benefit, service is denied to legitimate users. This is one of the most difficult attacks to thwart.

B. The defense for each Email security threat

1) The defense for Phishing attack is authentication. By using an authentication agent or digital certificates, you force the user to prove his or her identity. Through authentication you ensure that only trusted users can engage in sessions. [10]

2) The defense for the man in the middle and session hijacking attack is digital certificates or digital signatures. Both Parties of communication should proved to each other; that they know a secret that is known only to them. This Is usually done by digitally signing a message and sending it to the other party, also asking the other party to send a digitally signed message. [10]

3) The defense for eavesdropping attack is encryption using where only the authorized recipient will be able to decrypt.

4) The defense for dictionary attack is strong passwords. Passwords that are not common name,(like fist name, last name, or birthrate), words or references are harder to crack with a brute force attack such as a dictionary attack.

5) The defense for denial of service attack is authentication service filtering. By authenticating users on authenticated parties can send message.

6) The defense for Data diddling attack is a decrypted message digest. An encrypted mess digest records random segment of the original message so receiver recalculate the message digest, then compare it with the received message digest. If the message altered then encrypted again in its road, an encrypted message digest provides a method of authenticating the integrity of the data.

C. Email security protocols:

In order for making previous defenses for the main Email attacks ; to achieve the electronic mail security goals of confidentiality, message integrity, authentication, and non-repudiation from original, Mainly we have two protocols (PGP/MIME and S/MIME).

1) PGP

PGP is protocol provides a confidentiality and authentication service that can be used for electronic mail and file storage applications it had developed by the effort of a single person, Phil Zimmermann.

a) PGP functions:

Authentication:

In sender side, sender creates a message, then generate hash code of the message, which encrypted using the sender private key (the signature), and the result is concatenated with the message and compressed using ZIP.

In receiver side, receiver decompressed the message, sender's public key to decrypt and recover the hash code, Then receiver generates a new hash code, and compares it

with the decrypted hash code. If the two match, the message is accepted as authentic.

Confidentiality:

In sender side, sender generates a message and a random 128-bit number to be used as a session key for this message, then message is encrypted with the session key, the session key is encrypted with RSA using the recipient's public key and is pretended to the message.

In receiver side, receiver uses RSA with his private key to decrypt and recover the session key. Moreover, use the session key to decrypt the message.

message created using SHA-1(message digest), which encrypted using DSS or RSA with the sender's private key and included with the message.	
Confidentiality: Message encryption using CAST-128, IDEA, or 3DES, with a one-time session key generated by the sender. the session key is encrypted using Diffie-Hellman or RSA with the recipient's public key and included with the message.	CAST or IDEA or Three-key Triple DES with Diffie-Hellman or RSA

A. PGP message format as shown in figure (2).

TABLE I. CRYPTOGRAPHIC ALGORITHMS USED IN PGP.

Function	Requirement
Authentication: a hash code of a	DSS/SHA or RSA/SHA

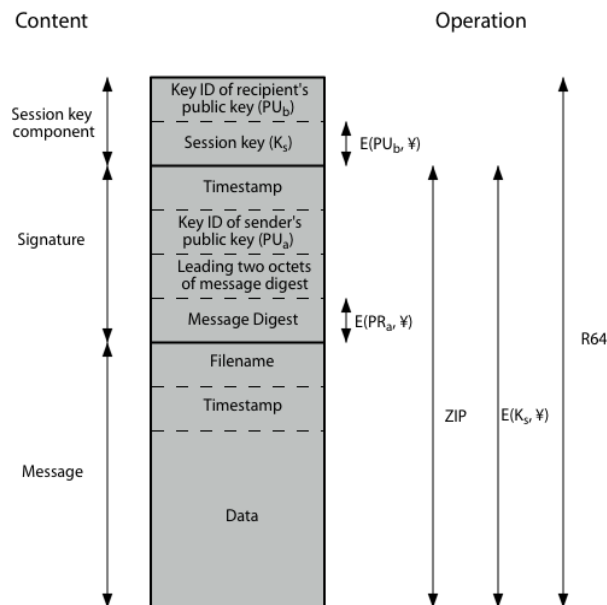


Fig.2 . General Format PGP Message.

b) PGP problems

PGP has several problems. "Key management is considered as a big challenge in PGP and PKI-based solutions in general. Public key cryptography requires the sender to obtain the receiver's public key beforehand, to be able to start any PGP encrypted communication (also it should be done in a secure way to prevent man-in-the-middle attacks). Moreover, there is still no practical secure approach to private key management; users should create a backup of their private key, store it in a safe place and be careful not to lose it, otherwise old encrypted emails cannot be decrypted anymore. Additionally, in case the private key is compromised, the attacker can trivially decrypt all the (old or new) encrypted emails. Therefore, a certificate revocation list (CRL) is required to facilitate the revocation of all compromised keys which also must be shared with all users". [14]

S/MIME is a protocol for adding cryptographic security utilities to e-mails. S/MIME requires no change in the sending and receiving MTAs process because this service can be added to the client software installed at sending and receiving clients. Basically its provide sender authentication, non-repudiation of sender, message integrity and message security using encryption and digital signatures.

a) MIME (review).

MIME is an extension to the RFC 5322 framework that is intended to address some of the problems and limitations of the use of Simple Mail Transfer Protocol (SMTP), MIME provided support for varying content types and multi-part messages.

MIME specification includes the following elements.

1. Five new message header fields are defined, which may be included in an RFC 5322 header as shown in figure (3).
2. A number of content formats are defined.
3. Transfer encodings are defined that enable the conversion of any content format into a form that is protected from alteration by the mail system.

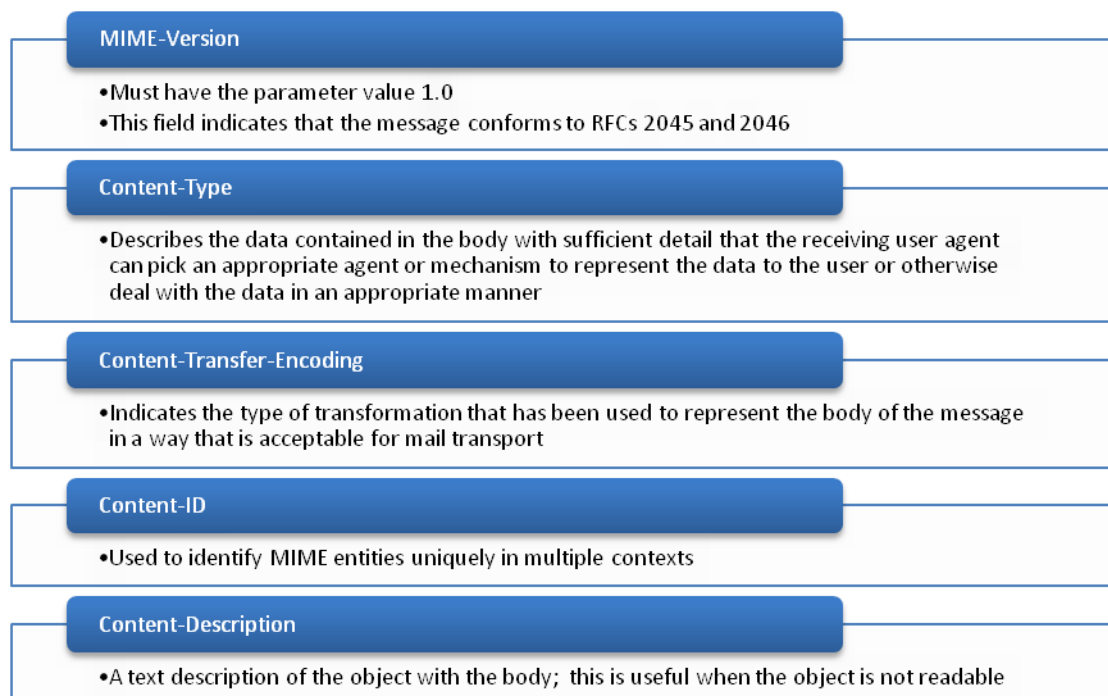


Fig.3 . MIME Headers

Example of MIME Message Structure:

```
From:Nathaniel Borenstein
<nsb@bellcore.com>
To: Nead Freed <ned@innosoft.com>
Subject: Sample message
MIME-Version: 1.0
Content-type:multipart/mixed;
boundary="simple boundary"
```

Hello. This section begins the actual message body,

b) S/MIME main functions.

- 1) Authentication and Confidentiality (Enveloped data): In sender side, sender prepare an envelopedData MIME entity by generate a pseudorandom session key, which encrypted with the receiver public RSA key, Encrypt the message content with the session key. This envelopedData is then encoded into base64. In receiver side, to recover the encrypted message, the receiver first strips off the base64 encoding, then the receiver's private key is used to recover the session key, Finally, the message content is decrypted with the session key.

- 2) Authentication (Signed data): In sender side, sender compute the message digest (hash function) of the content to be signed, Encrypt the message digest with the signer's private key, Prepare a block known as Signer Info that contains (signer's public key certificate, an identifier of the message digest algorithm, an identifier of the algorithm used to encrypt the message digest, and the encrypted message digest), then message and digest is encoded using base64. The Signer Info followed by the message constitute the signedData.

In receiver side, to recover the signed message and verify the signature, receiver strips off the base64 encoding, and then the signer's public key is used to decrypt the message digest. Receiver independently computes the message digest and compares it to the decrypted message digest to verify the signature

- 3) Signed and enveloped data: encrypted data may be signed and signed data or clear-signed data may be encrypted.

TABLE II. CRYPTOGRAPHIC ALGORITHMS USED IN S/MIME. [2]

Function	Requirement
Create a message digest to be used informing a digital signature	MUST support SHA-1
Encrypt message digest to form a digital signature.	Receiver SHOULD support MD5 for backward compatibility.
Encrypt session key for transmission with a message.	Sending and receiving agents SHOULD support Diffie-Hellman. Sending and receiving agents MUST support RSA encryption with key sizes 512 bits to 1024 bits.
Encrypt message for transmission with a one-time session key.	Sending and receiving agents MUST support encryption with tripleDES. Sending agents SHOULD support encryption with AES. Sending agents SHOULD support encryption with RC2/40.
Create a message authentication code.	Receiving agents MUST support HMAC with SHA-1. Sending agents SHOULD support HMAC with SHA-1.

A. S/MIME message format

"The MIME entity is prepared according to the normal rules for MIME message preparation. Then the MIME entity plus some security-related data, such as algorithm identifiers and certificates, are processed by S/MIME to produce what is known as a PKCS object. A PKCS object is then treated as message content and wrapped in MIME". [2]

Example of S/ MIME Message Structure:

```
From: Nathaniel Borenstein <nsb@bellcore.com>
To: Nead Freed <ned@innosoft.com>
Subject: Sample message
MIME-Version: 1.0
Content-Type: application/pkcs7-mime;
smime-type=signeddata;
name=smime.p7m
Content-Transfer-Encoding: base64
567GhIGfHfYT6ghyHhHUujpfyF4f8HHGTrfvhJh
jH776tbB9HG4VQbnj7
```

B. S/MIME problems.

Complexity of public key cryptography concept, and some user interface related usability problems of email clients supporting S/MIME (discussed are still barriers to S/MIME’s adoption. Since S/MIME is not broadly used due to the above mentioned problems, we do not discuss further S/MIME related proposals. [14]

III. CONCLUSION

To summarize the state of secure e-mail software, we can say that software exists now to establish trust between two individuals. Such software has actually been available for some time, but the quality and ease of use of available implementations has recently begun to improve. Software is available to secure MIME-based e-mail in a similar manner, although it is old as widespread and is mostly available commercially.

REFERENCES:

- [1] J Kurose. and K . Ross, "Computer Networking: A Top-Down Approach", 6th Edition, Addison-Wesley Longman, 2012.
- [2] Pfleeger and S. Lawrence Pfleeger. 2006. "Security in Computing" (4th Edition). Prentice Hall PTR, Upper Saddle River, NJ, USA.
- [3] W. Stallings, "Cryptography and Network Security Principles and Practice", Fifth Edition 2006.
- [4] Trend Micro Trend Labs Primer, "Trouble in your inbox. 5 Facts every small business should know about Electronic mail-based threats", Internet Security Threat Report, Volume 19, Oct 2012.
- [5] IDC, "Worldwide Messaging Security 2013 – 2017 Forecast and 2012 Vendor Shares", International Data Corporation Aug 2013.
- [6] McGuffin, "87-01-96 Security and Control of Electronic Mail by".
- [7] http://www.sans.org/?utm_source=web&utm_medium=textad&utm_content=generic_rr_pdf_interst1&utm_campaign=Reading_Room&ref=36923 , last visit 27/4/2014
- [8] [http://www.softlock.net/eSign-Electronic mail-Security](http://www.softlock.net/eSign-Electronic_mail-Security) last visit 27/4/2014
- [9] IDC, "Worldwide Security Software as a Service 2012 – 2016 Forecast: Delivering Security Through the Cloud" , International Data Corporation Dec 2012.
- [10] S. Setapa "Securing E-mail", SANS Institute Reading Room site, 2001.
- [11] P. Cooca, "Email Security Threats", SANS Institute Reading Room site, sep 2004.
- [12] http://infosecuritymag.techtarget.com/ss/0,295796,sid6_iss426_art874,00.html
- [13] M. Tariq Bandy, "EFFECTIVENESS and LIMITATIONS OF E-MAIL SECURITY PROTOCOLS", IJDPS, Vol.2, No.3, May 2011.
- [14] Pirouz, "SECURING EMAIL THROUGH ONLINE SOCIAL NETWORKS", August 2013

WSN for AIR Quality Monitoring in Annaba City

Mohamed FEZARI and Mohamed Seghir BOUMAZAI
Laboratory of Automatic and Signals Annaba
Badji Mokhtar Annaba University, Faculty of Engineering, BP:12, Annaba, 23000
Annaba, Algeria

Ahmed Al-DAHOUD , Ali Al-DAHOUD
Al-Zaytoonah University of Amman, Jordan , Faculty of IT, JORDAN
Amman, Jordan

Abstract: Wireless sensor networks (WSN) have been involved in different applications including monitoring many environmental phenomena such as air quality assessment, forest fire monitoring, flood rivers control.. In this paper, a WSN architecture where nodes are equipped with gas , temperature and dust sensors and Arduino-uno as microcontroller have been designed for air quality monitoring for some sensible area in Annaba City East of Algeria. The previous design included several units mainly: MSP430 Microcontroller, Gas and dust sensors, and the current regulator circuit. The new design is based on Arduino-Uno as micro-system. Comparing of normal gas levels for the clean air, the obtained results indicate that there is a big difference in the gas levels of both gases (LPG , NO₂ and CO) which obtained from the several tests. However, the acquired results for the air quality control in some areas in Annaba city show no risky situation to be considered for further actions. In this work we cover the field of Air quality monitoring electronic Nodes design and wireless transmission of fusion data. Then A GUI has been designed for simulation of the WSN in controlling the environment air Quality. Tests are encouraging; the flexibility off the shelf components and the ease of design facilitates the implementation of this system.

Keywords: WSN, AQM, air Quality Index, Environment Monitoring

1. INTRODUCTION

Wireless Sensor Networks (WSNs) technology [4] is in the front part of the investigation of the computer networks and it could be the next technologic market of with huge sum of money in investment. Sensor nodes can be fixed or mobile, they have limited processing power, storage, bandwidth, limited wireless transmission range and energy powered by battery. This limitation makes provision of the security in sensor networks not an easy task [4]. The availability of cheap, low power, and miniature embedded processors, radios, sensors, and actuators, often integrated on a single chip, is leading to the use of wireless communications and computing for interacting with the physical world in applications such as air quality control.

Sensor networks may consist of different types of sensors [5] such as seismic, low sampling rate magnetic, thermal, visual, infrared, acoustic and radar, which are able to monitor a wide variety of environment situations [5] such that: temperature, humidity, air quality, vehicular movement, lightning condition, soil makeup, noise levels, the presence or absence of certain kinds of animals or objects, mechanical stress levels on attached motors, and the current characteristics such as speed, direction, and size of an vehicle. A sensor node is made up of four basic components [5] as shown in Figure 2:

a of sensing unit, a processing unit, a transceiver unit and a power unit.

In this paper, we propose to use a WSN based microcontroller equipped with gas sensors have been actively used for air quality monitoring. The design included several units mainly: Arduino Microcontroller, MQ-2 Gas Sensors, and the current regulator circuit the paper si organized as follow: in second paragraph after introduction we define primary pollutants, in paragraph 3, we present the hardware proposition design with main components. In section 4, format and communication with the special sensor DHT11 is illustrated, then we finish the paper by presenting results ,discussion of simulation and conclusions in section 5 and 6 respectively.

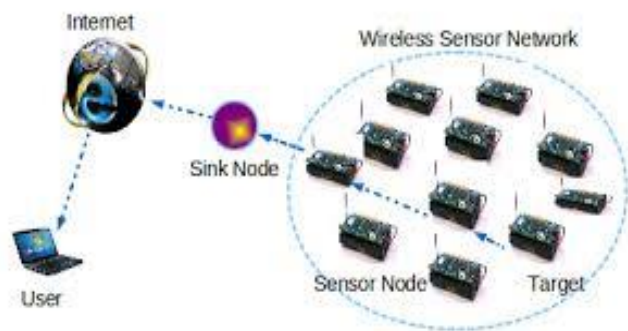


Fig. 1: The Hardware Design Schematic Diagram.

II. POLLUTANTS

Primary pollutants are those in which the substance emitted is itself hazardous. Some primary pollutants also produce other dangerous substances after undergoing chemical reactions in the atmosphere, and these are known as secondary pollutants. Primary pollutants include the following substances as mentioned in [15].

Particulates : This includes dust, smoke, aerosols and haze - any finely divided airborne solid material. Particulates are commonly generated by fires, motor vehicles, some industries (particularly road building, quarries and fossil fuel power stations) and various natural sources including volcanoes, plant and animal matter and dirt. Particulates are aesthetically displeasing, can irritate the eyes and cause respiratory problems. In recent years concerns have been raised about the possible health effects of 'fine' particulate matter (less than 10µm diameter). These have been shown to be associated with increases in hospitalization and even deaths from respiratory illnesses and heart disease.

Sulphur dioxide, SO₂ : Sulphur dioxide is often produced by the industrial processes which produce particulates, the primary sources of SO₂ being coal, fuel oil and diesel. Being a corrosive acidic gas, sulphur dioxide damages buildings and other materials, and can cause respiratory problems.

Carbon monoxide, CO : The commonest source of carbon monoxide is motor vehicle emissions, where it results from the combustion of petrol in the presence of insufficient oxygen. It is also a result of some fuel-consuming industries and domestic fires. Carbon monoxide is a colorless, odorless, highly toxic gas that displaces oxygen in human blood, causing oxygen deprivation.

The oxides of nitrogen, NO_x : NO_x refers to the mixture of nitric oxide (NO) and nitrogen dioxide (NO₂) formed by the oxidation of nitrogen during the combustion of air. The majority of NO_x is produced in motor vehicle emissions,

although other sources can have significant local impact. NO_x is a contributor to several secondary pollutants, and NO₂ is a respiratory irritant that can also corrode metals at high concentrations.

Benzene : Over the last few years leaded patrols have been phased out of use. However this has resulted in higher levels of benzene and other aromatics in the substitute unleaded petrol. Benzene breaks down quickly in the environment and is not stored in the tissues of plants or animals. However, it is still hazardous to humans at high levels as it can cause several diseases of the blood including leukemia (cancer of the white blood cells). Benzene monitoring programmes were started in New Zealand in 1994 and are continuing because the levels in some locations were found to be reasonably high.

Hydrogen sulphide, H₂S : Hydrogen sulphide is mainly associated with geothermal activity at Rotorua, where it is responsible for the 'rotten eggs' smell, but it is also formed from the anaerobic decomposition of many organic wastes and is a by-product of paper manufacture and leather tanning (see article). It is highly poisonous (more toxic than hydrogen cyanide), and because it initially anaesthetizes the sensory organs it can build up to high concentrations without warning and cause paralysis and then asphyxiation.

Fluorides : These have two main sources: the Comalco aluminum smelter and fertilizer works . Fluorides can have adverse effects on plants, and in some cases concentrate in the leaves so that animals eating the plants ingest significant quantities.

III. PROPOSED AIR MONITORING SYSTEM DESIGN

The complete system design is shown in figure 1, Hardware Design Schematic Diagram. The design is based on nodes and the architecture of the node contains the following major hardware components:

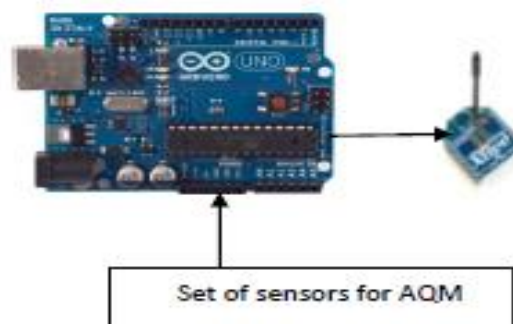


Fig 2.a: Sensor Node main Components

a. *Arduino Microcontroller* [1]: this is the core component of the design. Arduino is a flexible programmable hardware platform designed for fast Embedded Systems platform conception. Arduino's little, blue circuit board, mythically taking its name from a local pub in Italy, has in a very short time motivated a new generation of DIYers of all ages to make all manner of wild projects found anywhere from the hallowed grounds of our universities to the scorching desert sands of a particularly infamous yearly arts festival and just about everywhere in between.

Usually these Arduino-based projects require little to no programming skills or knowledge of electronics theory, and more often than not, this handiness is simply picked up along the way. In figure 2.b we can see the main components in the arduino-uno system board.



Fig 2.b: Arduino-uno system board

b. *MQ-2 GAS Sensor* [3] Breakout Board: MQ-2 is one of the series of semiconductor Gas Sensors that is used mainly for gas (such as CO) leak detection for houses, workshops, commercial building, Fire, Safety detection system as well as a gas leak alarm.

Resistance value of MQ-2 is difference to various kinds and various concentration gases. So, When using this components, sensitivity adjustment is very necessary. we recommend that you calibrate the detector for 1000ppm liquified petroleum gas <LPG>, or 1000ppm iso-butane <i-C4H10> concentration in air and use value of Load resistance that (RL) about 20 KΩ (5KΩ to 47 KΩ).

This sensor module utilizes an MQ-2 as the sensitive component and has a protection resistor and an adjustable resistor on board. The MQ-2 gas sensor is sensitive to LPG, i-butane, propane, methane, alcohol, Hydrogen and smoke. It could be used in gas leakage detecting equipments in family and industry. The resistance of the sensitive component changes as the concentration of the target gas changes.

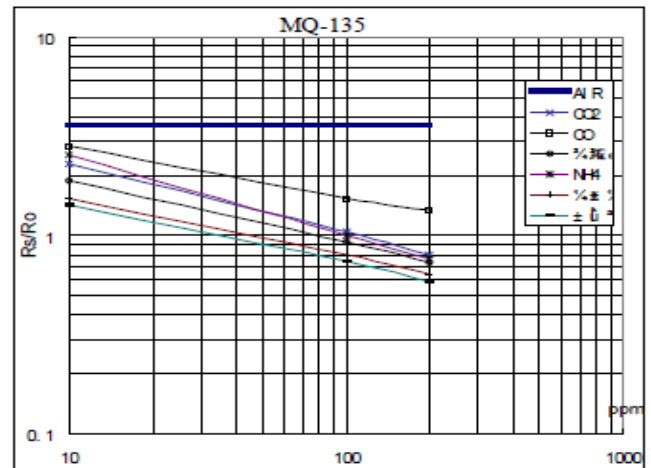


Fig 2.c: Front and rear image of the MQ-135 Gaz sensor; in Blue, potentiometer to control threshold and sensitivity characteristics of MQ-135 under T=22°C and H=65%, from Technical data MQ-135 sensor.

c Temperature and Humidity sensor

The DHT11, DHT21 and DHT22 are relative cheap sensors for measuring temperature and humidity. In reference [6] and [7] there is a description of library for reading both values from these sensors. we contacted the manufacturer to get the details of the differences between the two DHT sensors to build a lib that supports both. The DHT21/22 is quite similar to the DHT11 and has a greater accuracy (one decimal) and range (negative temperatures). The hardware pins and handshake are identical but it uses a different data format.

Communication and format for DHT11 : Single-bus data format is used for communication and synchronization between MCU and DHT11 sensor. One communication process is about 4ms.

Data consists of decimal and integral parts. A complete data transmission is **40bit**, and the sensor sends **higher data bit** first.

Data format: 8bit integral RH data + 8bit decimal RH data + 8bit integral T data + 8bit decimal T data + 8bit check sum. If the data transmission is right, the check-sum should be the last 8bit of "8bit integral RH data + 8bit decimal RH data + 8bit integral T data + 8bit decimal T data".

d *Resistance Circuitry:* Resistance value of MQ-2 is difference to various kinds and various concentration gases. So, When using this components, sensitivity adjustment is very necessary.

we recommend that you calibrate the detector for 1000 ppm liquified petroleum gas <LPG>, or 1000 ppm iso-butane <i-C4H10> concentration in air and use value of Load resistance that (RL) about 20 KΩ (5KΩ to 47 KΩ).

e. *ADC (analog-to-digital converter)*: is a device that converts a continuous quantity to a discrete digital number. Typically, an ADC is an electronic device that converts an input analog voltage (or current) to a digital number proportional to the magnitude of the voltage or current.

The conversion of analog signal and conditioning is performed by Arduino Uno Microcontroller.

f. *Light Emitting Diodes*: two LEDs used as indicators the Green one indicates the Battery level and the Red one indicates Gas concentration.

g. *Trasmission module* is based on Zigbee protocol, the Xbee modules are used in transmission and reception, they are also used as gateway node to provide data to the central unit.



Fig 2.c: Transmission Module Xbee

IV. DHT11 SENSOR PROGRAMMING AND PROTOCOLE

Single-bus data format is used for communication and synchronization between MCU and DHT11 sensor. One communication process is about 4ms.

Data consists of decimal and integral parts. A complete data transmission is **40bit**, and the sensor sends **higher data bit** first.

Data format: 8bit integral RH data + 8bit decimal RH data + 8bit integral T data + 8bit decimal T data + 8bit check sum. If the data transmission is right, the check-sum should be the last 8bit of "8bit integral RH data + 8bit decimal RH data + 8bit integral T data + 8bit decimal T data".

Source code for DHT11 sensor reading by Arduino uno:

in there code lines we illustrates part of the software to be included into the arduino uno memory.

```
#include "dht.h"
int dht::read11(uint8_t pin)
{
    // READ VALUES
    int rv = read(pin, DHTLIB_DHT11_WAKEUP);
    if (rv != DHTLIB_OK)
    {
```

```
        humidity = DHTLIB_INVALID_VALUE; // invalid
        value, or is NaN preferred?
        temperature = DHTLIB_INVALID_VALUE; // invalid
        value
        return rv;
    }

    // CONVERT AND STORE
    humidity = bits[0]; // bits[1] == 0;
    temperature = bits[2]; // bits[3] == 0;

    // TEST CHECKSUM
    // bits[1] && bits[3] both 0
    uint8_t sum = bits[0] + bits[2];
    if (bits[4] != sum) return DHTLIB_ERROR_CHECKSUM;

    return DHTLIB_OK;
}
```

V. RESULTS AND DISCUSSION

The proposed design were used to measure the air quality in several places inside the Annaba City and included different gases levels but focused mainly on measuring three main gases: Carbone Monoxide (CO) and Liquid Petroleum Gas (LPG) and NO2. A sample of obtained results from three different places : clean environment in Seraidi mountains at Annaba , Annaba city center where there is a crowd circulation and El-Hadjar region where a Metal-Steel production firm is installed at 10Km from Annaba city center, the results are shown in table 1.

Figure 3 and figure 4, are used to simulate AQM in region in Annaba city by acting on gas concentration levels. Simulation results: for simulation of WSN nodes, the area is divided into parts where each part can be controlled by a node, in this case the area is divided into 9 regions, and the transmission circuit is chosen so that it can provide the adjacent nodes with the information with minimum consumption of energy.

Scenario 1: by adjusting the sliders for CO, SO2 and NO2 Gas we obtained the Red color of the region , which illustrates by node 9 the values sensed: Co=206ppm SO2=160 ppm and NO2=200 ppm

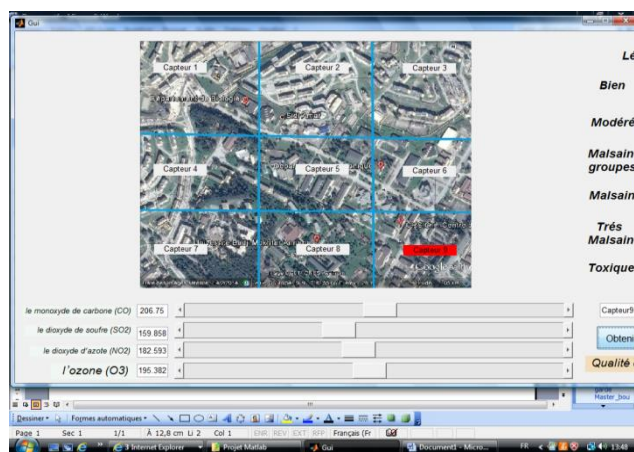


Fig 3: node 9 sensed $CO_2=206ppm$ $SO_2=160$ ppm and $NO_2=200$ ppm levels the central control unit

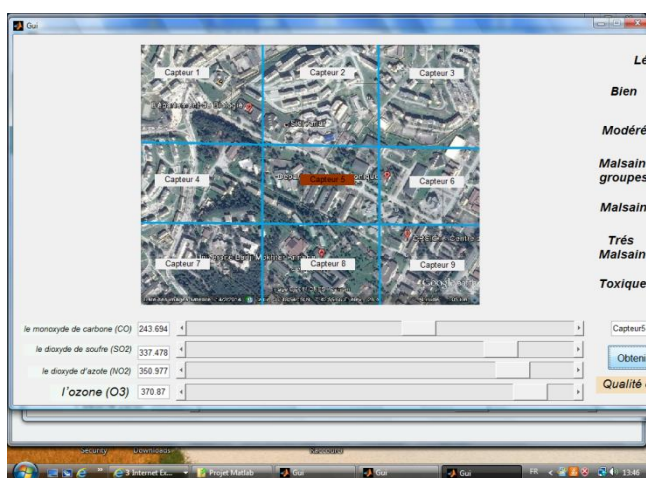


Fig. 4: Concentration of Carbon Monoxide Levels in Clean Air using WSN

Based on the normal gas levels of the clean air [2], the results indicate that there is a big difference in the gas levels of both gases (LPG and CO) which obtained from the several tests and circuit runs. However, the acquired results show no risky situation to be considered for further actions.

VI. CONCLUSIONS

AIR Quality monitoring System Design to assess the pollution of air in some parts of Annaba city using a micro-system, as a node in Wireless Sensor Network (WSN), is proposed in this article. WSN enhanced the process of monitoring many environmental phenomena such as the air pollution monitoring issue in proposed this paper. It provides a real-time information about the level of air pollution in different regions, as well as provides alerts in cases of drastic change in quality of air. Based on collected information, such data can then be used by the authorities to take prompt actions

such as evacuating people or sending emergency response team. The proposed design is enhanced by several ways such as: selecting adequacies' sensors, calibrating these sensors for gas detection, integrating them in a WSN system controlled by an Arduino-Uno, and finally transmission to the central unit using Xbee modules. A Graphic user interface has been presented in this work to simulate the effect of sensors on selected area. The results are interesting, improvements can be done: in providing a web service page that can provide these data to users, as well as more sophisticated sensors could be used such as MQ-135, MQ-136 and others.

Acknowledgements

Authors appreciate the support of LASA laboratory at Badji Mokhtar Annaba University (BMAU), faculty members of Environment department (BMAU) and Dean of Faculty of IT at Al-Zaytoonah University AMMAN, JORDAN.

REFERENCES

- [1] Mobile Air Quality Monitoring Network at www.isis.vanderbilt.edu/projects/maqumon.
- [2] S. Choi, N. Kim, H. Cha, and R. Ha " Micro Sensor Node for Air Pollutant Monitoring: Hardware and Software Issues ". Sensors 2009.
- [3]. Technical Data For MQ-2 Gas Sensor, Website <http://www.seedstudio.com/depot/datasheet/MQ-2.pdf>
- [4] Qasem Abu Al-Haija, "Toward Secure Non-Deterministic Distributed Wireless Sensor Network Using Probabilistic Key Management Approaches", Journal of Information Assurance and Security 6 (2011) 010-018.
- [5] I.F. Akyildiz, W. Su, Y. Sankarasubramaniam, E. Cayirci, " Wireless sensor networks: a survey", Computer Networks 38 (2002) 393-422, elsevier.
- [6] <http://playground.arduino.cc/Main/DHT11Lib>
- [7] <http://www.micro4you.com/files/sensor/DHT11.pdf>
- [8] Raja Vara Prasad Y. et al,"Real Time Wireless Air Pollution Monitoring System", ICTACT Journal on Communication Technology, June 2011.
- [9] M. Riley, " Programming Your Home: Automate with Arduino, Android, and Your Computer", The Pragmatic Programmers, 2012.
- [10] R.A. Roseline, Dr.P. Sumathi, "Local Clustering and Threshold Sensitive routing algorithm for Wireless Sensor Networks", in the IEEE sponsored International Conference on Devices Circuits and Systems(ICDCS'12), March 2012. (Available online at ieeexplore.com).
- [11]. Honicky, R.; Brewer, E.A.; Paulos, E.; White, R. N-smarts: networked suite of mobile atmospheric real-time sensors. In *Proceedings of the 2nd ACM SIGCOMM Workshop on Networked Systems for Developing Regions*, Seattle, WA, USA; ACM: Seattle, WA, USA, 2008.
- [12]. Volgyesi, P.; Nadas, A.; Koutsoukos, X.; Ledeczi, A. Air Quality Monitoring with SensorMap. In *Proceedings of the 7th International Conference on Information Processing in Sensor Networks*, St. Louis, MO, USA; IEEE Computer Society: St. Louis, MO, USA, 2008.
- [13]. So, S.; Koushanfar, F.; Kosterev, A.; Tittel, F. LaserSPECKs: laser SPECTroscopic trace-gas sensor networks - sensor integration and applications. In *Proceedings of the 6th International Conference on Information Processing in Sensor Networks*, Cambridge, MA, USA; IEEE Computer Society: Cambridge, MA, USA, 2007.
- [14] Sharma, A.; Golubchik, L.; Govindan, R. On the Prevalence of Sensor Faults in Real-World Deployments. In *4th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks*, San Diego, CA, USA; IEEE Computer Society: San Diego, CA, USA, 2007.

[15] Saitas, Jeff; *Ground-Level Ozone*; [Online] Available <http://www.tnrce.state.tx.us/air/monops/ozoneinfo.html>; February 20, 1997.

[16] Xintaras, Charlie and Perry, Mike; *Agency for Toxic Substances and Disease Registry*; [Online] Available <http://atsdrl.atsdr.cdc.gov:8080/toxfaq.html>; February 20, 1997

TABLE 1: SITUATION OF AIR POLLUTION IN THREE DIFFERENT AREAS IN ANNABA CITY

Clean Air Co	Clean Air NO2	Clean Air LPG	Center Co	Center NO2	Center LPG	Metal- Steel Co	Metal- Steel NO2	Metal- Steel LPG
0.05	1	2.05	20	16	75	35	56	25
0.8	2.5	3.5	26	13	86.9	34	57	24
0.75	0.8	2.7	24	17	87.4	36.7	58	38
0.48	0.8	1.9	26.78	15.68	80.6	40.58	55.8	26.25
0.87	2.4	2.9	27.58	19	76	32.78	50.15	27.8
0.79	1.7	3.04	29.15	20	79	31.99	52	30
0.61	1.5	2.9	30.15		78.95	32.58	53	30.5

Complex Adaptive WSNs for Polluted Environment Monitoring

Amjad RATTROUT

Department of Computer Science
Arab American University
Jenin, Palestine
Amjad.ratttrout@aaup.edu

Hamzah Hijawi

Department of Computer Science
Arab American University
Jenin, Palestine
hamzah.hijawi@gmail.com

Abstract—The development of low cost, low power, multi-functional sensor nodes created a new form of networking applicable to a variety of fields. The goal is to create low cost, low power consumption and web accessed wireless sensor nodes in order to be used for environment monitoring. A wireless sensor network consists of large number of sensor nodes. One of these nodes is used as a gateway to be connected to the internet; this node is called a base station or a sink node. The capabilities of the gateway node will be larger than the other nodes in the system. However it will be directly connected to the internet via Wi-Fi or Ethernet adapter. Moreover a web server may be installed on this node; this will enable the system to be directly accessed any time from any location. The low power consumption is strongly needed to be used in A wireless sensor networks; this because the wireless sensor nodes almost powered by batteries. In order to achieve this goal an intelligent hardware and intelligent software should be implemented carefully to adapt the environment changes and do some adjustments and calibrations on the system.

Keywords—WSN, Multi Agents, Complexity, Co-evolution

I. INTRODUCTION

Wireless sensor networks are emerging technologies of the past few years; they involve a large number of small nodes. Each node senses environmental changes and report them to other nodes over flexible network architecture. They are varied in their applications and areas [1] [2]. The main components of a wireless sensor node are microcontroller, transceiver, and sensors. Microcontroller processes data collected by the sensors, controls other components in the node and performing power consumption management. Transceiver transmits/ receives data from/ to other nodes in the network. Sensors have the sensing tasks. Wireless nodes are powered by batteries.

The main challenges are to produce low cost, low power consumption and web accessed nodes [3]; generally, the transceiver consumes the largest amount of power. Therefore, it is an advantageous to send data only when it is required. This requires intelligent agents loaded to the sensor nodes which can convolute and make decisions according to the system status. Additionally, it is important to minimize the power consumed by the sensor itself. Therefore, the hardware should be designed to allow the microcontroller to judiciously control

power consumed by the transceiver and controlling the sensing frequency; however if the sensed phenomena is not critical, the microcontroller must be programmed to read the sensors only when a reasonable changes occur in the environment [4].

Wireless sensors networks are autonomous systems with severe energy and processing power limitation and constrain, however end nodes have limited reliability. In such conditions, self-organizing, energy efficient, fault tolerant and adaptive algorithms are required to be used in WSN [1].

Generally the main tasks of a wireless sensor node are to interact with it is environment (sensing) and reporting sensed data to the sink node. The sensing capabilities of different environment phenomenon emerge new applications for WSN; this includes: Environmental monitoring, health care systems, military, educations and smart home buildings [5]. WSN simplify the systems managements and helps building a smart world.

WSN systems can be deployed with different topologies; the most common are Bus, Tree, Ring and Star. In Bus topology all nodes connected via single link and when a node wants to communicate with another node, it sends a broadcast

message, all nodes connected to the bus will receive the message and only the destination node will process the message. In Tree topology the system is divided into levels, however nodes only communicate with their parents. In Ring topology every node has only two neighbors, in this topology the message is propagated to each neighbor until received by the destination node. In Star topology the sink or main node is logically located at the center point of the system; it will have a direct connection to each node within the system and nodes communicating by passing messages through the sink node [6]. Star topology will be used to build the clusters in our system.

In this research, we will focus in the software intelligence part and build smart agents to control the system. The agents will use some mathematical calculations to decide which nodes are redundant and putting them in the sleep mode. Nodes in the sleep mode will wake up periodically and check if they still redundant. The rest of this research is organized as follow. Related work is given in section 2. Section 3 gives an overview about the system architecture used. And the problem is defined and discussed in section 4. Section 5 discusses the proposed method to implement the system using the multi agents. In section 6 we propose two mechanisms to reduce the power consumption in the system. Section 7 is the simulation part and finally conclusions are discussed in section 8

II. RELATED WORK

The problem of intelligent implementation of wireless sensor network had been studied extensively in that past few years [1] [3] [4]. Most of researches focus on the small part of the system [7] [8] and almost none of them provided a complete system solution. Wireless sensor networks threads were studied in normal client server architectures [2] and multi agents approaches were discussed in [4] [8] [9]. The complexity and adaptation in multi agents systems were extensively studied in Web systems [10]. However in [11] they discussed the complexity in data collection, data aggregation and data selection over the tree model in which the sink node which is the first level in the tree is connected with two nodes to form the second level and so on. The conative WSN was studied in [12], however in this model the network can make decisions and actions based on the condition of the environment and the current system status to achieve its goals. BDI model which is used to implement the intelligent agents was introduced in [14]; they described the general architecture of this model and defined the belief, desire and intention in the wireless sensor node, using this model the agents will be aware of their environment and can cooperate together to achieve the system goals. An example was introduced in [16] in which the wireless sensor network was used to implement an intelligent transportation system.

In this research we will provide a complete solution for implementation intelligent wireless sensor network using multi agent approach and will show the system components, types of agents and their interaction.

III. SYSTEM ARCHETICTURE

The system will be built based on a cluster topology in which for each cluster there is a cluster head node which is responsible for coordination and data aggregation from other member nodes. Cluster head nodes have direct connection to this sink node. Sink node considered as a gate way of the system. However it is the connection between our environment and the external word. Sink node and cluster head nodes have more processing power and longer life time from other nodes in the system because they are considered as key nodes and have more extra work to do. Sink node is connected to the external word using internet and the collected data will be stored in external data base.

For simplification we considered that the system has three clusters and nodes have direct connection to their cluster head.

Fig. 1. System architecture

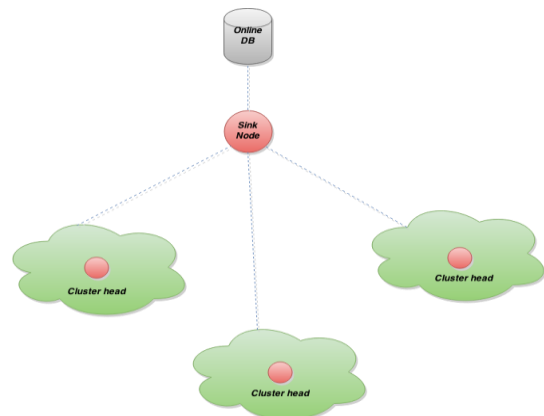
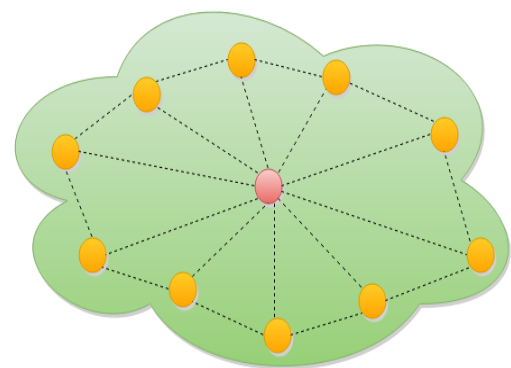


Fig. 2. Cluster architecture



IV. PROBLEM DEFINITION

Wireless sensor network is an open complex adaptive system, in the following sections we will describe the most

important characteristics of WSN and we will propose a method using intelligent agents to resolve those issues.

A. Environment Description

The main application of WSN is environment monitoring, and in most cases it is the natural environment [2]. WSN gives humans clear observation and detailed studies about the surrounding environment by collecting data using sensors. Applications involved in the environment monitoring are developed from data we get from sensors to information after processing the data and finally to knowledge the human gains about the real status of his environment. Application scenario explained in this research is related to collecting data from a polluted environment, this type of applications requires collecting data much as possible. The collected data is processed to give clear information or can acts as input to another system in order to do some actions on the environment. For simplification measuring CO₂ concentration will be used as an example, however any other example is applicable.

B. Complexity of WSN

Wireless sensor network is a complex system; complexity can be seen in many fields. However Data collection, data aggregation, processing the gathered information, time complexity, messages complexity and energy cost complexity are the most complex elements in WSN [10]. Different types of components and large number of nodes make the complexity in systems like WSN. However in data collection, we need to decide which data we need to collect from sensor nodes and how to aggregate the raw data, this includes all kinds of queries and routing algorithms used to deliver data to the base station node. Processing data at sensor nodes requires complex operations; however do we need to process all the data or just pick up some samples. Reduction energy consumptions in sensor nodes almost the most important feature in WSN, however many algorithms are used to decide when and how to stop sensing, increase/ decrease sampling rate, even sleep the node entirely. To make the system simpler and function well, a kind of decisions making needed to be implemented as will be shown in next sections.

C. Adaptation and Co-evolution

WSN always interacts with its environment; however it should adapt to the environment changes and convolutes to increase the system life time [10]. Kinds of adaptations include decrease the sampling rate when the recent sensed data has the same values. Another kind of adaption in this system is the election processes to elect a new cluster head node when the old node wants to die due to low remaining battery power. The aim is to make an intelligent system which able to make decisions and actions based on the conditions of the environment. Intelligence when applied to WSN can make the system behaves better and will increase the performance [12]. The system should be aware of the sensed data, when and how to forward it, this will result in better power and bandwidth management over the overall network. WSN implementation using agents will enable the system to have a high level of knowledge about the environment and the type of information being exchanged, this will help in achieving the system

objectives by making the network aware of and adapts to the application requirements and the environment in which it is deployed.

V. MULTI AGENT APPROACH IN WSN

Traditional methods such client-server architecture used in WSN in which each node senses the environment independently from other nodes and propagate the data to it is neighbors until reaching the sink node can work with small size networks, however this approach has many issues; what if the system get larger and larger and we need to cover more areas, all nodes are active and they flood the sensed data in the network, this will put a heavy load on the other nodes because they have additional task to propagate other nodes data and as a result of that network performance will decrease. Even more, what about the sink node capability to handle and process this huge amount of data coming at the same time? Bottleneck will appear at some points causes unexpected faults [4]. Using multi agents in the sensor networks can solve the scalability, transparency and performance issues discussed previously. Redundancy elimination and saving power consumption also can be achieved with a good implementation of intelligent agents. However for each specific task in the system a specific type of agents will be created [8].

Belief, desire and intention (BDI) model is the most popular model for implementing intelligent agents. Multi agents may exist in a single complex system and almost they have common task which performed by each of agents independently from others. Agent may communicate with each other via messaging protocols [13]. However the goal is to increase the level of abstraction by using belief, desire and intentions instead of instruct the agent exactly what to do and how to do. Intelligent agent should be able to decide and make decisions according to the system variables.

A. Agents Types and Functions

Taking in consideration that following tags are used:

- S_NODE: Tag to identify the sink node.
- H_i_NODE: Tag to identify head cluster node, where i is the cluster number.
- N_{ij}_NODE: Tag to identify each node within its cluster, where j is the node number.

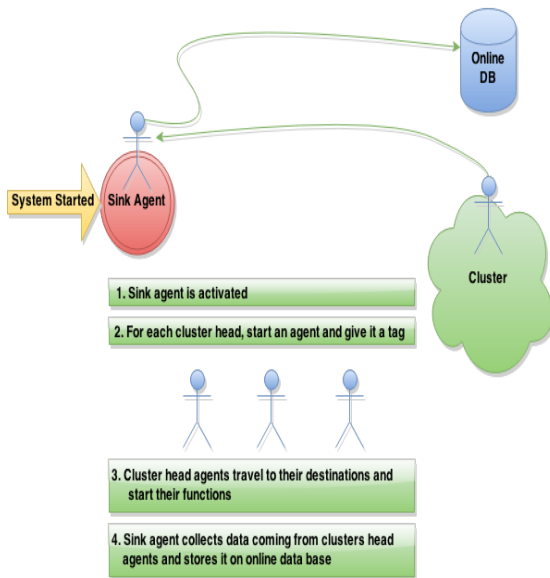
System is divided into multiple agents, each agent is specialized only in one function and he will die after complete his task. Following are the intelligent agents used in our system:

- Sink agent (Main system agent).
- Head clusters agents.
- Environment sensing agents.
- Control agents.
- Tagging agents.
- Data aggregation agents.

Sink Agent (Main Cluster Agent)

When the system runs for the first time, the main agent will be activated. Simply main agent will be the brain of the system. It will initiate tagging and main cluster agents, in addition to collect the aggregated data and stores it on online data base.

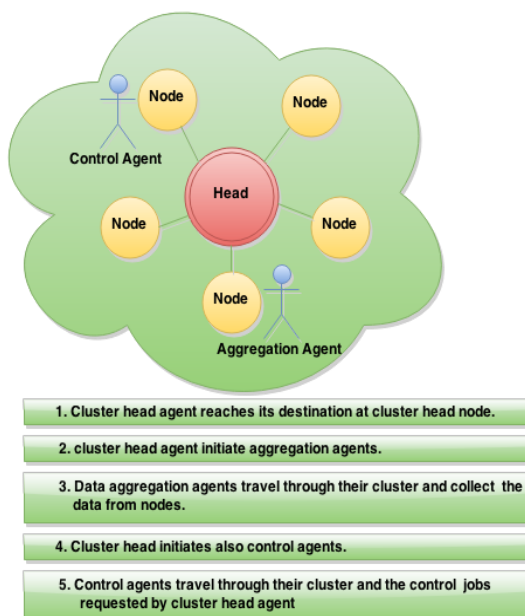
Fig. 3. Sink agent flow chart



Head Clusters Agents

Head cluster agents are created at sink node and travel to their destinations at cluster head nodes. All other agents in the cluster are created in the cluster head and they travel to sensor nodes in order to collect data or do some controls within their cluster.

Fig. 4. Head cluster agent flow chart



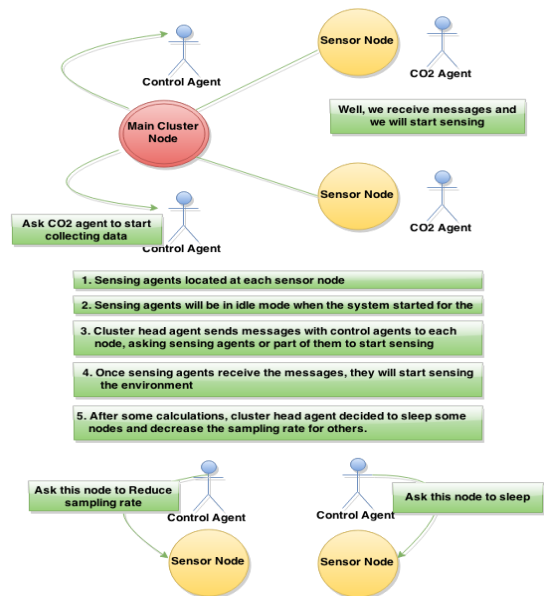
Sensing Agents

Sensing agents lie on each sensor node and they have only one function which is to read an environment variable, however for each variable there is a special agent, this includes agents responsible for reading CO2 concentrations, agents to read VOCs, etc. Sensing agents started their work upon request from the head cluster agent through the control agents.

Control Agents

Control agents are initiated from head cluster agents, their job is to control the cluster nodes as requested from head cluster agent, this includes checking redundant nodes and put them in the sleep mode and stop/start sensing some environment variables.

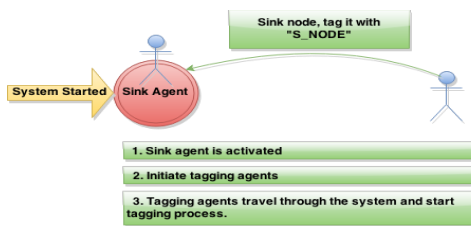
Fig. 5. Control and sink agents flow chart



Tagging Agents

Tagging agents are initiated at sink node using the sink agent and also at the cluster head nodes using cluster head agents. However tagging agents created at sink node are responsible for tagging the sink node and the cluster head nodes. Whereas tagging agents created at clusters head are responsible for tagging each node in their cluster. Tagging agents have additional task which is to increase the revision number of each sensor node visited by aggregate agents during data aggregation process.

Fig. 6. Tagging agents flow chart



Data aggregation agents

Data aggregation agents travel through the network and collect data from sensor nodes. However each sensor node has small database. When the aggregation agent reaches a sensor node, it will concatenate that table to its current table and empty the local table.

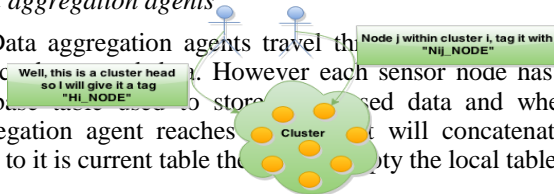
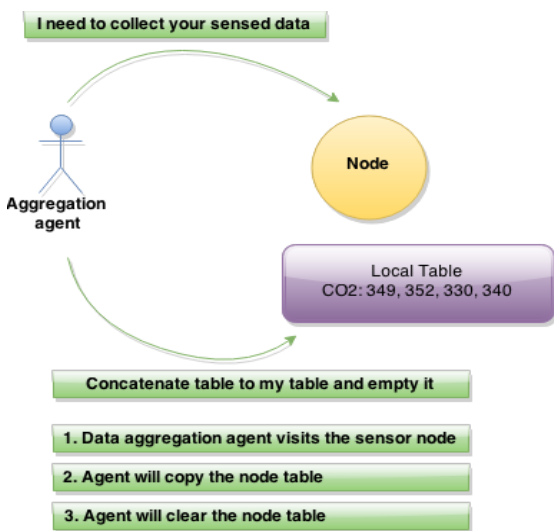


Fig. 7. Data aggregation agents flow chart



B. Agents Interaction

Most applications in WSN use multiple mobile agents to collect, share and process information about large scale environment. Information collected by agents need to be reported to the external users through the system gateway which is the sink node. Collecting information by agents and send it back to the sink node only when it is necessary will take less power than each agent sends his information alone [8] [9]. So how the system internally works? How the agents interact and cooperate to achieve the system tasks? First, when the main agent start running, it will have information about the number of clusters in the network and the clusters head IDs, actually only this information is needed for the main agent to start his work. So it will create one cluster head agent for each cluster. Cluster head agents travel through the network until each of them reaches its destination, however the agent knows his destination by comparing the cluster head ID he carries

with the node ID, if it is the same then the agent will be activated on the main cluster node and start managing his cluster.

Cluster head agent only needs the information about what environment variables should be collected by the sensor nodes, so for each variable it will create sensing agent which will travel to the sensor nodes. Sensing agents will start sensing tasks immediately when they reach the nodes, however each agent has small table to store the sequence of collected data. At some point the cluster head agent decided to collect the sensed data, so he will create multiple aggregation agents which will travel through the nodes. When the aggregation agent reaches the sensor node it will append the node tables to its table. Aggregate agent will continue with this process until he reaches his maximum payload size, at this point he will return back to the cluster head agent which will process the collected data. Aggregation agent will die after this step.

But how aggregation agent knows if another agent already visited this sensor node and collected the data? The proposed idea is to give the aggregation agent revision number when the cluster head agent create them, when those agents visit the nodes for the first time they will ask the tagging agent to tag the node with same revision number. The revision number is an incremented value, so for the next time the cluster head agent wants to collect the data; he will create a new aggregation agents with incremented revision number. When the aggregation agent reaches the node he will compare his revision number with sensor node revision and will copy the node tables only when it is revision number is greater than the node revision, and then it will increment the node revision number.

Another type of agents created by cluster head agent is the control agents, we need this agent in order to control, manage and detect faults in the cluster. Suppose for some reasons – will be described in later sections – main agent decided to put some nodes in the sleep mode, so it will create a control agent which will travel to those nodes and asks them to be in the sleep mode for a specific period of time, moreover control agents may ask some sensing agents to stop collecting data.

C. Building Trust between Agents

Building a trust in multi agent system is a very important step that needs to be taken in consideration. The needs of trust appears in the complex adaptive systems such as WSN because of the openness and diversity of the system components, however agents may frequently enter and leave the system which results in changing of the system structure. In such conditions it will be very difficult for the agents to build a confidence communication between each other. It is worth to mention that in most multi agent systems it is impossible to build 100% trust between agents, trust level is relative to the system goals. However an example, when the aggregate agent visits the sensor node to aggregate the data stored in its local database, sensor node needs to trust the aggregate agent and be sure it is part of the system. To build such a trust first the cluster head agent should encapsulate the aggregate agent and add the cluster ID to its header, when the aggregate agent

reaches the sensor node, sensing agent will compare its cluster ID with the cluster ID provided from the aggregate agent and will provide its table only if they have the same cluster ID. Another model which can be used is to use a trust agent which acts as a broker; however when the aggregate agent tries to communicate with the sensing agent in order to collect its data it will be checked by the trust agent which will decide if it is trusted or not according to its tag and policies. However these are just examples to illustrate the trust between aggregate and sensing agents, many other models can be used [18].

VI. POWER SAVING MECHANISMS

There are many methods and mechanisms to reduce power consumption in sensor networks systems. Intelligent hardware design and implementation can almost achieve a good percentage of this goal. However intelligent agents can also contribute to achieve this goal. However using BDI model, the following methods are proposed to decrease the power consumption in WSNs.

A. Put Redundant Nodes in Sleep Mode

In order to estimate the value provided by a sensor node many mathematical expressions can be used to calculate the duplication in the system. In our system of monitoring polluted environment an intelligent agent will be implemented in the cluster head node, it will check if there is a redundancy in the cluster and put the redundant node in the sleep mode [3].

How it works? Taking in consideration that our system is built based on cluster topology in which for each cluster there will be a head node and there will be a direct connection between the cluster head and each of the other nodes on that cluster. Using the variance for calculations estimations:

- A cluster head agent will be assigned to each cluster.
- It will communicate with other agents in its cluster and ask them to calculate for example CO₂ concentration at that point.
- Each node will send its calculated value back to the cluster head agent.
- The cluster head agent will save temporary the CO₂ concentration with node ID for each sensor node.
- The cluster head agent will calculate the average of CO₂ concentration based on its value and the other nodes values.
- If the difference between the average value and the real value is within the error range, then put that node in a sleep mode.
- Nodes in a sleep mode will wake up periodically and check again if they still redundant

B. Reduce Sampling Rate

Sampling rate in wireless sensor network depends on the application, if we are dealing with critical environment and we

are concern about the recent data; high sampling rate should be used. However low sampling rate can be used for less critical applications such as environment monitoring which in most cases don't do any action on the environment. In this research which studies the polluted environment we use the BDI model to decrease the sampling rate because in such applications the data doesn't change often, so there is no need to have high sampling rate which consumes battery power [17].

Reduction of sampling rate is the control agent task which is described in the following steps:

- Cluster head agent who has the most recent collected data sees that the data values almost the same so he decides to reduce the sampling rate.
- Cluster head agent creates multiple control agents.
- Control agents travel through their cluster and ask the sensing agents to reduce the sampling rate.
- Control agents will die after finishing their task.
- Same procedure if the cluster head agent decided to increase the sample rate.

C. Ask Sensor to Stop Sensing the Environment

Based on our model in which each wireless node is responsible for sensing more than one variable. Sometimes one of the sensors reports duplicated data or we are concern about the sensed data over a specific period of time. In this cases control agent interacts with sensing agent and asks it to stop sensing.

- Sink agent receives a request from network admin to stop sensing CO₂ concentration.
- Sink agent creates multiple control agents loaded with a message to stop sensing CO₂.
- Control agents travel through cluster head nodes and deliver the message.
- Each cluster head agent creates multiple control agents which travel through the cluster.
- Once the control agent reaches the sensing node, it contacts with agent responsible for CO₂ sensing.
- Control agent will die and the sensing agent will stop sensing.

D. Ask nodes to Stop Processing Collected Data

At the critical point, when the overall remaining power in the system is very low. One approach to increase the life time is by doing the calculations and processing in the external world, this simply happens by asking the sensing nodes to forward the raw data collected from the environment.

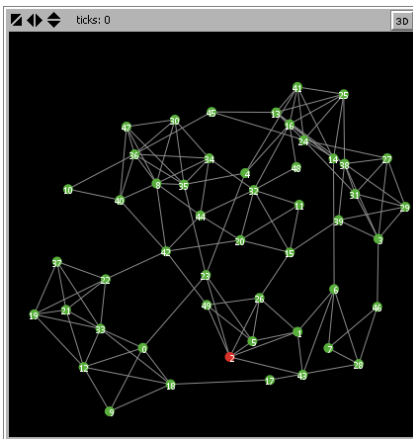
- Sink agent creates multiple control agents loaded with a message to stop processing data.
- Control agents travel through cluster head nodes and deliver the message.

- Once the cluster head agent receives the message, it will create multiple control agents with the same message.
- Control agents travel within their clusters and ask the sensing agents to stop processing the collected data.
- Sensing agents will forward the raw data as it is collected from the environment.
- Control agents will die.

VII. SIMULATION

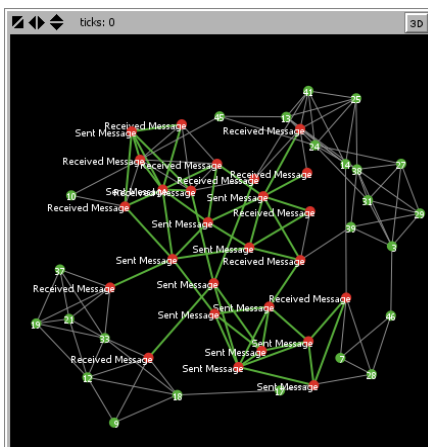
Netlog simulator is used to implement multi agents in wireless sensor network. From simulation we can conclude the benefits of using multi agents over normal client server architecture in complex systems such WSN. However client-server model consumes systems resources and communication bandwidth.

Fig. 8. Initial cluster with 50 sensor nodes



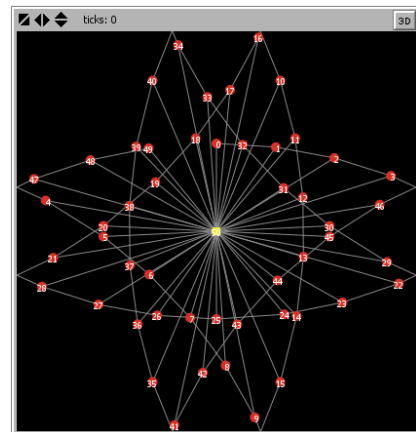
In normal client server architecture in order to send the message from the source to its destination, the source node starts flooding the message to all its neighbor. Once the message reaches the neighbors they will also resend it, this process continues until the message reaches all the nodes including the cluster head node.

Fig. 9. Flood a message in client server architecture



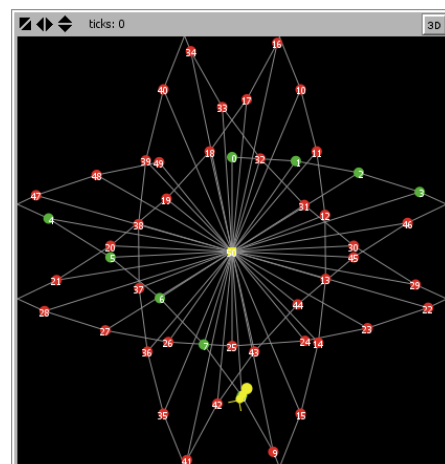
Next we show how to aggregate the data from sensor nodes using the aggregation agents with a cluster based topology, however the communication bandwidth is reserved and the data aggregate is more efficient.

Fig. 10. Initial cluster based on star topology with 50 nodes



Now to aggregate the data stored at sensor nodes and for simplicity one aggregation agent is generated and it visits the nodes one by one and collecting the data stored in their local tables. However when its payload reach the maximum size it will return to the cluster head node and provides the data then dies. A new aggregation agent is generated and it will continue with the same process until the data is collected from all sensor nodes.

Fig. 11. Aggregating data using the aggregate agent



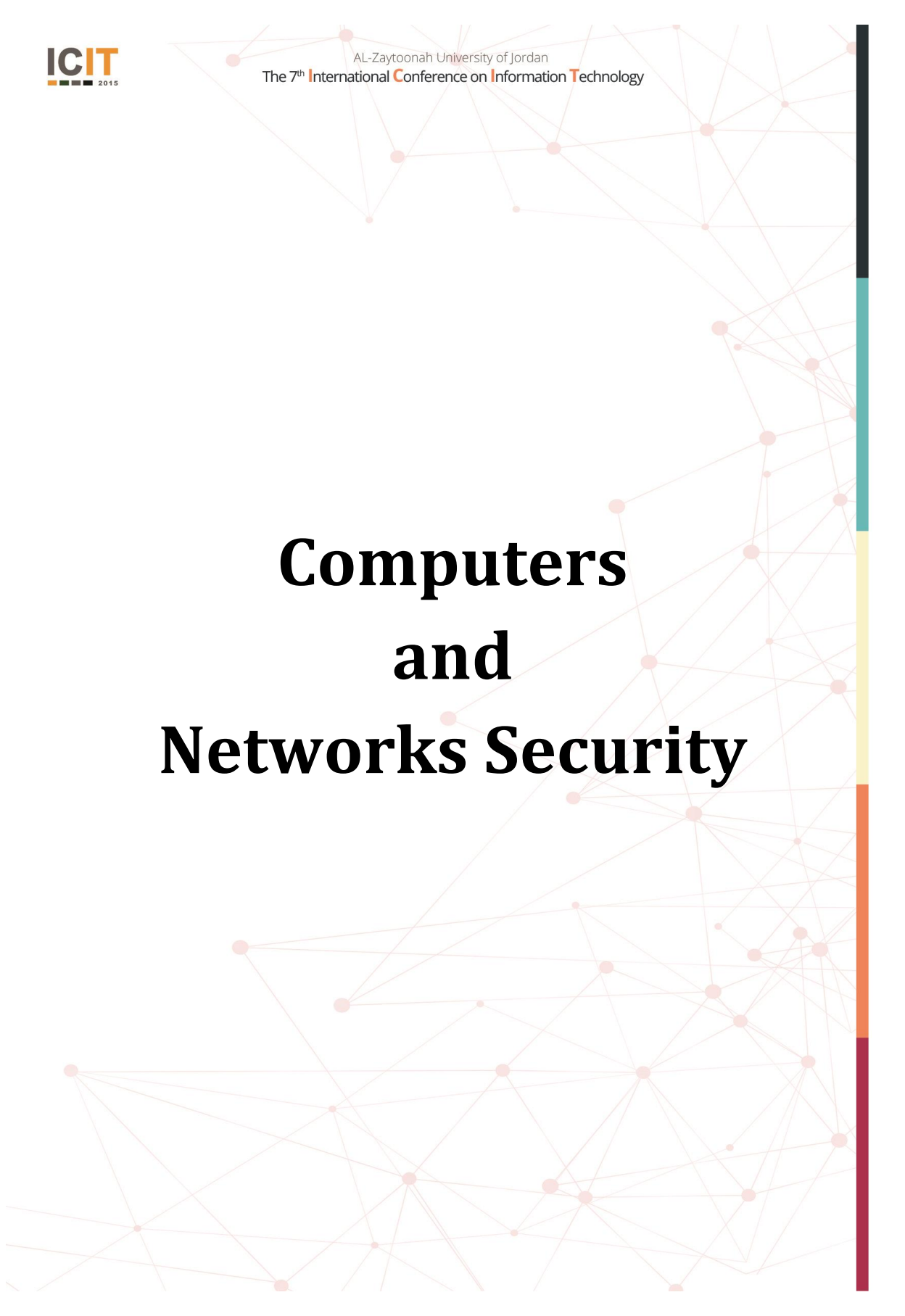
VIII. CONCLUSION

In this research we have proposed the multi agents paradigm for wireless sensor networks implementation, however agents are used for control, sensing and aggregating data. Complexity, scalability and adaptation of wireless sensor networks were studied in this research and a suitable solution using intelligent agents were proposed to address those aspects of such systems. In WSN where we have a large number of sensing nodes distributed over a specific area almost there are many redundant nodes which collect the same data, this type of redundancy is needed in WSN and we can consider it as a good redundancy if we have the control over the network, however we showed how the intelligent agents can do the control jobs over their clusters and put the redundant nodes in the sleep mode and wake them up when they are not still redundant. Moreover agents can interact and communicate with each other, travels through the network to achieve their tasks. We proposed four methods for power saving in this system and showed how to do that using intelligent agents. Almost agents will be the brain of the system and will do whatever needed. Comparing multi agents approach with normal models used in WSN implementation, the complexity decreases significantly because of agents specialty, scalability increases and the system is easy to adapts.

REFERENCES

- [1] Gungor, Vehbi C., Bin Lu, and Gerhard P. Hancke. "Opportunities and challenges of wireless sensor networks in smart grid." *Industrial Electronics, IEEE Transactions on* 57, no. 10 (2010): 3557-3564.
- [2] Alam, Sahabul, and Debashis De. "Analysis of Security Threats in Wireless Sensor Network." *arXiv preprint arXiv:1406.0298* (2014).
- [3] Morris, Alexis, Paolo Giorgini, and Sameh Abdel-Naby. "Simulating bdi-based wireless sensor networks." In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 02*, pp. 78-81. IEEE Computer Society, 2009.
- [4] Fortino, Giancarlo, Stefano Galzarano, Raffaele Gravina, and Antonio Guerrieri. "Agent-based Development of Wireless Sensor Network Applications." In *WOA*, pp. 123-132. 2011.
- [5] Kuorilehto, Mauri, Marko Hännikäinen, and Timo D. Hämäläinen. "A survey of application distribution in wireless sensor networks." *EURASIP Journal on Wireless Communications and Networking* 2005, no. 5 (2005): 774-788.
- [6] Sharma, Divya, Sandeep Verma, and Kanika Sharma. "Network Topologies in Wireless Sensor Networks: A Review 1." (2013).
- [7] Krishnamachari, Bhaskar, and S. Sitharama Iyengar. "Efficient and fault-tolerant feature extraction in wireless sensor networks." In *Information Processing in Sensor Networks*, pp. 488-501. Springer Berlin Heidelberg, 2003.
- [8] Haghighi, Mo. "Cooperative Task Allocation in Utility-Based Clustered Wireless Sensor Networks." *International Journal of Information and Electronics Engineering*, Vol. 3, No. 6, November 2013
- [9] Tynan, Richard, David Marsh, Donal O'kane, and Gregory MP O'Hare. "Intelligent agents for wireless sensor networks." In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pp. 1179-1180. ACM, 2005.
- [10] Rupert, Maya, Amjad Rattrout, and Salima Hassas. "The web from a complex adaptive systems perspective." *Journal of Computer and System Sciences* 74, no. 2 (2008): 133-145.
- [11] Li, Mo, Yajun Wang, and Yu Wang. "Complexity of data collection, aggregation, and selection for wireless sensor networks." *Computers, IEEE Transactions on* 60, no. 3 (2011): 386-399.
- [12] Haque, Md Alimul, Md Faizanuddin, and N. K. Singh. "A Study of Cognitive Wireless Sensor Networks: Taxonomy of Attacks and Countermeasures." (2012).
- [13] Shin, Jaewon, AM-C. So, and Leonidas Guibas. "Supporting group communication among interacting agents in wireless sensor networks." In *Wireless Communications and Networking Conference, 2005 IEEE*, vol. 4, pp. 2375-2380. IEEE, 2005.
- [14] Sangualagi, Prashant, A. V. Sutagundar, S. S. Manvi, and Vidya S. Bennur. "BDI Agents for Information Fusion in Wireless Sensor Networks." *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)* 1, no. 7 (2012): pp-144.
- [15] Chen, Min, Sergio Gonzalez, Yan Zhang, and Victor CM Leung. "Multi-agent itinerary planning for wireless sensor networks." In *Quality of Service in Heterogeneous Networks*, pp. 584-597. Springer Berlin Heidelberg, 2009.
- [16] Wang, Hao. "Wireless sensor networks for an extended city intelligent transportation system." *International Journal of Advancements in Computing Technology* 3, no. 5 (2011): 300-307.
- [17] Alippi, Cesare, Giuseppe Anastasi, Mario Di Francesco, and Manuel Roveri. "An adaptive sampling algorithm for effective energy management in wireless sensor networks with energy-hungry sensors." *Instrumentation and Measurement, IEEE Transactions on* 59, no. 2 (2010): 335-344.
- [18] Momani, Mohammad, and Subhash Challa. "Survey of trust models in different network domains." *arXiv preprint arXiv:1010.0168* (2010).

Computers and Networks Security



Enhancing Intrusion Detection System (IDS) by Using Honeybee Concepts and Framework

Ghassan Ahmed Ali

College of Computer Science and Information System
Najran University
Najran, Kingdom of Saudi Arabia
gaabdulhabeb@nu.edu.sa

Abstract— Intrusion Detection System has been studied for more than ten years. Though Artificial Intelligence (AI) techniques have been integrated to improve IDS but the success is still far from satisfaction. Thus, we believe a new strategy to improve IDS is badly needed. One of the solutions is by imitating the honeybee colony that can successfully protect their colony. In fact, the honeybee colony system and problems are quite similar with network system, and the way the bees protecting their colony can also be considered similar with the IDS in the network system. In this paper, we investigated the honeybee colony system as well as their detection system to get improvement methods for IDS engine in order to enhance IDS system for a better network defense. The adaptation of the honeybee protection and defense system itself is a new knowledge that can help other systems such as antivirus, antimalware, or even defense system to imitate the AI techniques in performing their functions. We train the proposed system with different types of attacks data and model different types of attack signatures. The performance of the proposed IDS is evaluated using NSL-KDD data set. The experiments show that the performance of the proposed model can detect novel intrusions and reduce false alarms.

Keywords— honeybee; intrusion detection; system Security

I. INTRODUCTION

The accuracy of detecting intrusion is directly depending on the accuracy of classification which is the first layer of IDS. Poor classification will result in the occurrence of intrusion and false alert [1]. A classification method is very important to obtain effective countermeasure against the intrusions.

The ability to recognize and detect intrusion is critical to the maintenance of the integrity of social insect colonies. Therefore, many researches take steps toward supporting computer security by understanding the methods underlying social insects' behavior system which face the same problems and see how there system works.

The crossover between the behavior of social insects and computer science can be declared as “. . . any attempt to design algorithms or distributed problem-solving devices inspired by the collective behavior of social insect colonies and other animal societies . . .” by Bonabeau et al. [2]. From studying how social insects perform tasks, we figure out such model to be used as a basis of development, either by enhancing the model or by adding non biological features to the model. The most important is the applicability of the model. The mimicry in all details is kind of exaggeration; to a certain extent, the similarity that it deduces to be useful should be the most concern.

The intelligent behaviors of honeybee have been developed to different models and methods which are applied for solving various types of problems. In the literature survey some studies modeled the honeybee foraging or finding home to be used in optimization problem [2]. Other works have proposed models based on the marriage behavior of honeybee [4]. From these

models there being extracted many features were being utilized by engineering and computer science [5].

Despite the strength of security system of honeybee behavior in nature (such as guarding, perception, information flowing, nest policy and rules, etc), however, it is still "raw material" in computer sciences application. Previous research in biology has shown that honeybee guard made very few errors in accepting nest-mates and rejecting non-nest-mate [6]. In addition, Honey bees use an early-warning system to detect threats and defend the nest [7]. The multilayer protection in honeybee colony and the diversity of defenses can be viewed as a distributed detection system. All these features in the behavior of the honeybees can be a construction of a novel security model to develop the accuracy of IDS.

Honeybees in nature survive in difficult environments, different levels of threats to security. These threats motivate bees to be able to detect and respond quickly on any aggressive acts that may attack the colony [8]. This paper focuses on how bees solve such security problems regarding the detection to crossover directly to IDS. This can be achieved through the use of the approach and architecture that are based on honeybee mechanisms. The investigation of this approach yield new insight into computer sciences.

This paper investigates a new method in the direction of construct a significant decision to accept or reject the incoming packets based on packet characteristic that each packet posses, in addition to get accurate decision to accept valid packets and reject intruder. This development for packet classification will improve robustness and accurate detection.

II. THE HONEYBEE GUARD APPROACH AND DETECTOR COMPONENT

The mechanisms ‘D-present/U-absent’ which is used to match the characteristics among individuals and rejecting non-group members in nature is proposed as a model by [9]. This model assumes that recognition system of nest-mates is detecting either the presence or the absence of the characteristics they carry.

In this paper, the methods Undesirable-Absent (UA) and Desirable-Present (DP) that honeybee guard uses in nature in order to filter the incomer are applied to IDS. Undesirable-Absent (UA) calculates the undesirable characteristics that found in a receiving packet and compares it with the internal characteristics template. In order to apply the idea of undesirable-absent in the domain of intrusion detection, we need to determine the characteristics that will represent the malicious or attacks (the non-nestmate in real honeybee). For this reason, the dataset collected by DARPA 1999 and preprocessed for the KDD '99 competition have several relevant features that can be used as characteristics for the attack properties. Neural network will be trained to recognize the characteristics of attacks in order to classify these characteristics as undesirable characteristics during the testing phase.

The Desirable-Present detector compares between the characteristics of the forwarded packet and its "template" which contains the desirable characteristics of an accepted packet. The Desirable-Present detector built of normal data. The normal of "10% KDD" Cup 1999 dataset, which is free from attacks used to train the Desirable-Present detector in order to recognize the desirable characteristics of the incoming connection records. The advantage is to aim the Desirable-Present detector to detect new types of intrusions; as unexpected intrusions are deviating from normal network.

After preprocessing the data and training the neural network, the task would be to determine whether the test data belongs to normal or abnormal based on the features of connection records from a given new test data. The result of this learning process is a neural network which is capable of detecting anomalies in the traffic during the testing phase "corrected KDD".

The proposed IDS is divided into three main modules. Practically, each module is implemented to perform a designated intrusion detection task. Moreover, the generality of the detector is ensured by the standard data representation schemes for input/output adopted by the constituent modules.

III. STRUCTURE OF THE PROPOSED IDS

The core components of the detector modules consist of a set of soft computing classifier to have the ability to detect both well-known and novel intrusions. Figure 1 shows general

structure of the proposed system. The description and the interactions of the main modules are as following:

- Training Data Processing: A file called "Training Set" is input to the IDS. The file contains network data from the KDD Cup 1999 intrusion detection data set. Each row of the file contains an instance of the data, and each column represent unique attributes. The data also can be presented directly from the live network dumped from any sniffer. The task handled by the data processing module is to normalise or cleanses the given dataset for the data mining. Since the performance of any IDS not only depends on output of the IDS but also on input traffic

- Data mining: This module represents the data mining techniques, and uses the training data to train the system. The attributes of the trained data mining are stored for later use, during the testing. The output of the data mining module is a text file containing the parameters and the weight of the learnt neural network. This approach has the advantage of being able to automatically retrain intrusion detection models on different input data that include new types of attacks. The training processes are further explained in the coming sections.

- IDS Testing: In this phase, the data mining will be validated to ensure its usefulness. In order to prove that the proposed system is not only successful on the training set, a separate test set with new data is used to test the system. This data also comes from the KDD Cup 1999 set. Typically, network based IDS process system activities based on network data and make a decision to evaluate the probability of action of these data to decide whether these activities are normal or intrusion.

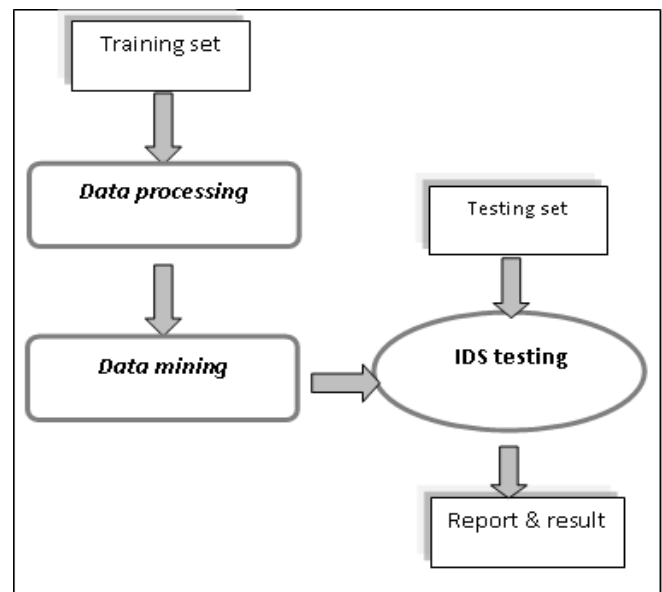


Figure 1: Summary of the Proposed IDS Structure.

In this study, we use Artificial Intelligence (AI) techniques in order to take the advantages of the new approach to improve the IDS. According to [10], the concept of using AI to solve the two IDS problems is very efficient. The generalization of AI makes possible decreases of false alarms as well as increases the accuracy of an intrusion detection process.

One of the important requirements for the technique to support the proposed approach is the ability of learning. Beside that, this technique is supposed to distinguish different characteristics after some level of training. Thus the neural network has been chosen to be the main component of the model because of the many features that neural network poses such as the ability of learning, generalizing attributes even with noisy data, and the capability of classifying patterns effectively. These features can be further used to improve detection and reduce false alarms in the intrusion detection system.

After the training phase, the neural network will be able to make the distinction between both normal and anomalous and then within anomalous between different attack classes. Once the neural network is trained, it can be used to classify new data sets whose input/output associations are similar to those that characterize the training data set.

1. THE TRAINING COMPONENTS PART

The objective of the training part is to train the neural network such that it becomes perceptive and sensitized to the specified dataset. The training components train the neural network such that the internal structure or topology of the given dataset. At first, the data set read by the initialization function. Then, the weights of the neural network are generated by the Bees Algorithm training. From the data file and the parameters given by the user, the initialization function will provide the user with random values as weights. The summary of training process illustrated in Figure 2. Once the network is trained, it can be used to classify new data sets whose input/output associations are similar to those that characterize the training data set.

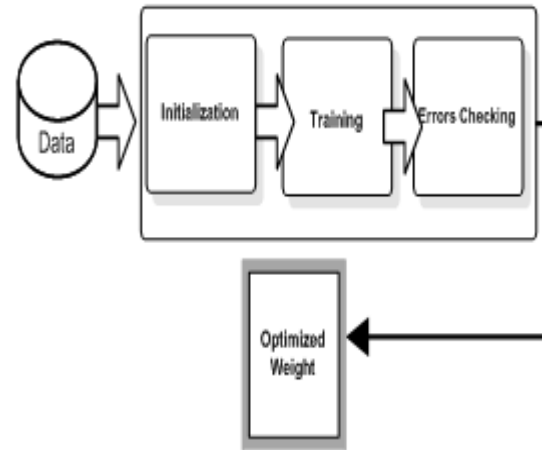


Figure 2: Neural Network Training

A. Neural Network Training

In the proposed work, the problem and data clearly indicate that the neural network learning is the supervised learning type. The training data task consists of T input-output (vector-valued) data pairs as following:

The Neural Network (NN) consists of a set of neurons or nodes which are interconnected with each other. According to [11], each neuron in the network is able to receive input signals, to process them and to send an output signal. Moreover, each neuron is connected at least with one neuron, and each connection is evaluated by a real number, called the weight coefficient, that reflects the degree of importance of the given connection in the neural network.

$$u(n) = (x_1^0(n), \dots, x_k^0(n))^t, d(n) = (d_1^{k+1}(n), \dots, d_t^{k+1}(n))^t \dots 1$$

where n denotes training instance. The output of the neural network is a function of synaptic weights W and input values x , i.e., $Y = f(x, W)$. The i th neuron can be written as equation 2

$$Y_i = f_i(\sum_{j=1}^n w_{ij} x_j + \theta_i) \dots 2$$

Where y_i is the output of the node, x_j is the j th input to the node, w_{ij} is the connection weight between the node and input x_j , θ_i is the threshold (or bias) of the node, and f_i is the node transfer function.

$$E(w(t)) = \frac{1}{n} \sum_{j=1}^n \sum_{k=1}^K (d_k - o_k)^2 \dots 3$$

where, $E(w(t))$ is the error at the t th iteration; $w(t)$, the weights in the connections at the t th iteration; d_k , the desired output node; o_k , the actual value of the k th output node; K , the number of output nodes; n , the number of patterns.

Record Type	No. of Patch	No. of Detection Records		FN	FP
		UA	DP		
NSL-KDD	1st_Patch= 1000 records	620	330	20	30
	2nd_Patch= 1000 records	407	593	0	0
	3rd_Patch= 1000 records	498	489	7	6
	4th_Patch= 1000 records	795	200	0	5
	5th_Patch= 1000 records	962	38	0	0
	6th_Patch= 1000 records	169	820	4	7
	7th_Patch= 1000 records	823	177	0	0
	8th_Patch= 1000 records	338	659	1	2
	9th_Patch= 1000 records	572	421	3	4
	10th_Patch= 1000 records	619	380	0	1
Overall		5803	4107	35	55
The Overall Rate		99.1%		0.35 %	0.55%

2. EVALUATION CRITERIA

Detection rate and a false positive rate are two main performance indicators. The false positive rate especially is critical to the performance of an intrusion detection system as a small difference of the false positive rate may translate into high number false alarms compared to the actual number of real alarms [1]. In most of the situations, it is not the ability of identifying attacks but rather its ability of suppressing false alarms that limit the performance of an intrusion detection system. The two major indications of performance are illustrated below:

$$DR = \frac{\text{detected intrusion samples}}{\text{total number of samples}} \quad (4)$$

$$FPR = \frac{\text{normal samples incorrectly classified as intrusion}}{\text{total number of samples}} \quad (5)$$

3. USING NSL-KDD_2009 TO TEST THE PROPOSED APPROACH

The new data set, NSL-KDD as suggested by [12], which consists of selected records of the complete KDD data set is using to test the proposed approach. The data set is publicly

available for researchers and has advantages over the original KDD data set.

The new dataset can be applied as an effective benchmark data set to help researchers to compare different intrusion detection methods [13]. The generated data sets, KDDTrain+ and KDDTest+, included 125,973 and 22,544 records, respectively. A 20% subset of the KDDTrain+.txt file is used for training the proposed IDS system whereas a subset of the KDDTest+.txt file is used for the testing phase. Table 2 shows the overall results on the NSL-KDD dataset.

Table 2: Experimental Result in Test NSL-KDD Dataset.

Table 2 illustrates the high performance of the proposed IDS. It shows the higher detection rate 99.1% and a low False Positive Rate 0.55% and False Negative Rate 0.35% of the system performance. The results obtained in this test demonstrate clearly the benefit of the proposed approach on the NSL-KDD dataset. More specifically, it can be observed that Undesirable-Absent detector is indeed capable of detecting more than half of the intrusions either new or old whilst the task of Desirable-Present detector is efficiently demonstrated; it is obvious that most of the undetectable intrusions by Undesirable-Absent are detected by Desirable-Present detector. In practice, the Desirable-Present detector is more sensitive and restrictive if found any variation from normal data. The combined of Undesirable-Absent and Desirable-Present detectors in proposed approach leads to get high detection rate and low false alarm.

Result from Specific Population Testing

In this experiment, the performance measure of proposed IDS is tested with specific population testing. The attacks in the data set fall into four main categories: DoS, R2L, U2R, and PROBE. In order to demonstrate the abilities of detecting different kinds of intrusions, the training data and testing data cover all intrusion categories. Totally, 1,200 attack data and 1,000 normal data were prepared for training and another set of 1,200 attack instances and 1,000 normal data were selected as the testing data. The attack population data are selected according to the measure attack categories and have the same approximate distribution as the KDD data set. The selected data records are illustrated in Table 3 below.

Attack Category	Attack Name	Records	Total
Normal		1000	1000
DoS	Neptune	155	517
DoS	Smurf	174	
DoS	Back	92	
DoS	Land	40	
DoS	Apache2	33	
DoS	Teardrop	23	
Probe	Ipsweep	129	

Probe	Nmap	59	
Probe	Portssweep	77	
Probe	Satan	44	
Probe	Mscan	36	
Probe	Saint	24	
U2R	buffer_overflow	82	217
U2R	sqlattack	79	
U2R	Perl	8	
U2R	Xterm	22	
U2R	Rootkit	26	
R2L	guess_passwd	41	97
R2L	Imap	2	
R2L	ftp_write	22	
R2L	Phf	20	
R2L	Sendmail	12	

Table 3: Experimental Result from Initial Population Testing

In the experiment, the performance measure of *Undesirable-Absent* and *Desirable-Present* detector are carried out solely on the selected data subset from the corrected file of the KDD'99 dataset which contains test data with corrected labels and other attacks examples from 10% KDD. The primarily results show that it's possible to increase the detection rate and reduce false alerts. Each method in honeybee approach has a good performance in identifying intrusion patterns and detects attacks. Table 4 shows the experiment results. The results show that *Undesirable-Absent* & *Desirable-Present* detectors have high *Detection Rate* and low *False Positive* even with small data set

Record Type	No. of Records	No. of Detection Records			
		UA	DP	DR %	False Alarm
Normal	1000	17	963	963/1000= 96%	17/1000=1.7%(FP)
Probe	369	202	165	367/369= 99%	2/369=0.5%(FN)
DoS	517	328	188	516/517= 99.8%	1/517=0.19%(FN)
U2R	217	82	134	216/217= 99%	1/217=0.46%(FN)
R2L	97	22	73	95/97= 98%	2/97= 2.1%(FN)

Table 4: Experimental Result from Selected Population Testing

The proposed approach demonstrates better performances in the most number of attacks categories and less false alarm. Based on the results that shown in previous Tables, it can be seen that the proposed approach has a good performance for detecting intrusion in computer networks. Moreover, the overall result of the detection of old and new attacks in different classes are high.

IV. CONCLUSION

The focus of this paper was to demonstrate how productive the crossover between biology and computer science can be. The detection system in honeybee, which keeps the colony safe, was the basis frame of the research to improve the effectiveness of IDS. The new approach is used to improve the IDSs at the detector level to distinguish between the innocuous and intruders using the way that honeybee is used in nature. Characterizing the incoming packets to support detection was significant. Characterization methods have ranged using trained neural network that it becomes perceptive and sensitized to detect intrusions.

To examine the feasibility of our approach, we conducted several experiments. The experimental results demonstrate that the proposed approach can improve the detection deficiency issue by reducing the false alerts and increasing the detection accuracy.

REFERENCES

- [1] Jan N.Y., Lin S.C., Tseng S.S. and Lin N.P. (2009), A decision support system for constructing an alert classification model, *Expert Systems with Applications* 36, pp. 11145–11155.
- [2] E. Bonabeau, M. Dorigo, G. Theraulaz, "Swarm Intelligence: From Natural to Artificial Intelligence", NY: Oxford University Press, NewYork, 1999.
- [3] Ali, G.A., Jantan, A., Ali, A.: Honeybee-Based Model to Detect Intrusion. In: Park, J.H., Chen, H.-H., Atiquzzaman, M., Lee, C., Kim, T.-h., Yeo, S.-S. (eds.) ISA 2009. LNCS, vol. 5576, pp. 598–607. Springer, Heidelberg (2009)
- [4] Yang C, Jie Chen J, Tu X (2007a) Algorithm of fast marriage in honey bees optimization and convergence analysis. In: IEEE international conference on automation and logistics, Jinan, pp 1794–1799
- [5] Ali, Ghassan Ahmed and Jantan, Aman: A New Approach Based on Honeybee to Improve Intrusion Detection System Using Neural Network and Bees Algorithm. Springer Berlin Heidelberg. 2011. http://dx.doi.org/10.1007/978-3-642-22203-0_65
- [6] Grüter C, Kärcher MH, Ratnieks FLW (2011). The Natural History of Nest Defence in a Stingless Bee, *Tetragonisca angustula* (Latreille) (Hymenoptera: Apidae), with Two Distinct Types of Entrance Guards. *Neotrop. entomol.* <http://dx.doi.org/10.1590/S1519-566X2011000100008>.
- [7] Breed, D., Guzmán-Novoa, E., Hunt, G.J.: Defensive behavior of honey bees: organization, genetics, and comparisons with other Bees. *Annual Review of Entomology* 49, 271–298 (2004)
- [8] Couvillon, M. J., & Ratnieks, F. L. W. (2007). Odour transfer in stingless bee marmelada (*Friesocomelitta varia*) demonstrates that

- entrance guards use an “undesirable-absent” recognition system. *Behavioral Ecology and Sociobiology*, 62(7), 1099-1105. Springer.
- [9] Sherman P.W., Reeve H.K., Pfennig D.W. Recognition systems. In *Behavioural ecology: an evolutionary approach* Krebs J.R., Davies N.B. 1997pp. 69–96. Eds. Oxford, UK:Blackwell Science.
- [10] Servin A. and Kudenko D. (2008). Multi-agent reinforcement learning for intrusion detection, *lecture notes in computer science*, vol. 4865; 2008. p. 211–23.
- [11] Daniel Svozil, Vladimir Kvasnicka, Jiri Pospichal (1997). Introduction to multi-layer feed-forward neural networks, *Chemometrics and Intelligent Laboratory Systems*, Volume 39, Issue 1, November, Pages 43-62, ISSN 0169-7439, DOI: 10.1016/S0169-7439(97)00061-0.
- [12] Tavallae M., Bagheri E., Lu W., and Ghorbani A. (2009). A Detailed Analysis of the KDD CUP 99 Data Set. Submitted to Second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA).
- [13] Ghorbani, Ali A., Lu, Wei, Tavallae, Mahbod (2009). *Network Intrusion Detection and Prevention*, Springer US. Doi: 10.1007/978-0-387-88771-5_7

On The Improvement of the Tri-Way Pixel Value Differencing Steganography Algorithm

Nada Mahmoud Aboueata, Sara Yaqoob Al-Rasbi, Wafa Ahmed Al-Jaal, Jihad Al-Ja'am
Department of Computer Science and Engineering
Qatar University
Doha, Qatar
{nada.aboueata,sara.alrasbi,wa095710,jaam}@qu.edu.qa

Abstract—Steganography consists of hiding information into digital files so that they cannot be noticeable to human vision. These files include texts, images, audios, videos and protocols. In this work we consider only gray-scale images as they are the most used digital covers in steganography. The most challenging problem consists of finding the right pixels of the image to embed or hide the maximum amount of information without deteriorating its quality. Research in this area is still at the premature phase even though several steganography image-based algorithms have been proposed recently. In this paper, we study the well-known tri-way pixel value differencing algorithm (TPVD) aiming at improving its performance. We select randomly the starting pixel of the pixel-pair combinations to hide information rather than starting always with the first pixel as done in the TPVD regular behavior algorithm. Our first experiment shows that a slight improvement can be obtained with these random selections in preserving the quality of the stego-image (i.e., the image holding the information) and maximizing the amount of hidden information. We study also the encryption of the sensible information that should be embedded into the cover image using the AES algorithm. These randomness behavior along with the encryption technique render the retrieval of the hidden information very hard once the image is spotted as suspicious to be a stego-image and the hidden information are attempted to be extracted. Our second experiment shows that the encryption of the same amount of information may deteriorate the quality of the stego-image and makes it somehow perceptible for human vision and also vulnerable to stego-analysis techniques.

Keywords—*information security; steganography; secret communication; tri-way algorithm*

I. INTRODUCTION

Exchanging information over computer networks the Internet is a challenging task due to the possible attacks that may occur during the transmission phase. Several encryption algorithms have been proposed to secure the information and made them illegible once detected by an illegal interceptor. However, encrypted information can be easily vulnerable to analysis and then decryption attempts. Another technique is proposed to secure transmitted secret or sensible information. It consists of embedding the information into a cover digital image and make it imperceptible to human vision. This technique is known as steganography and the cover image that hides the information is called a stego-image. This means of covert communication can be used in commercials and military applications. An image consists of a set of different numbers representing intensities in different areas of an image. This number-based representation constitutes a grid and the points are called pixels. Each pixel in a gray-scale image is represented by 8 bits and can have 256 different intensities. Several image-based steganography algorithms have been proposed recently aiming at maximising the amount of information to be hidden and preserve also the image quality [12,13]. However, they require major improvements [15]. The

PVD algorithm is one of the popular approaches used in steganography. It consists of hiding information using the difference of two consecutive pixel-values. The stego-images obtained from the PVD algorithm and its derivatives can be easily detected by the difference histogram analysis techniques as they follow one regular direction in the embedding phase [15,16,17]. Luo et al. [15] have proposed a more secure steganography algorithm which consists of splitting the image into blocks and randomly rotating them. The resulting image is divided into units of three pixels where the middle one is selected to embed the information. Although the authors shown an improvement in the embedding phase, it can be considered as regular as it starts always with same pixel. Chang et al. [1] have proposed an efficient steganography algorithm called the tri-way pixel value differencing algorithm (TPVD). This algorithm consists of using always the first possible pixel-bit during the embedding phase of the information into the cover image. This selection is thought to preserve always the quality of the stego-image. In this paper we study the consideration of random pixels as starting pixels in the embedding of the information. We show that a slight improvement can be obtained with this technique. We study also the encryption of the information that need to be

embedded into the cover image and we show how the quality of the stego-image is affected.

The rest of the paper is organized as follows: in section 2, we detail the TPVD algorithm. In section 3 we present our method which considers random-based combination of pixels in the embedding phase. In section 4 we show the results of our experiments and finally in section 5 we conclude the paper.

II. THE TPVD ALGORITHM

The TPVD algorithm is proposed as a significant improvement of the PVD algorithm. Both techniques split the cover image into a sequence of 2x2 blocks of pixels. They use then the difference of each pixels-pair to determine the number of bits that could be embedded. They assume human vision can easily observe changes in gray values of smooth area of a stego-image, but they are unable to notice relatively larger changes at the edges areas [1,15]. In order to determine the smoothness properties of the stego-image, the difference between every pixel-pair is calculated. To find the numbers of bits that should be embedded in each pixels pair, a vector ranged from 0 to 255 is built. The range of gray values is divided into smaller ranges as follows: [0-7, 8-15, 16-31, 32-63, 64-127, 128-255] where each region is defined by a lower bound (Li) and an upper bound (Ui). The absolute value of the difference for each pixel-pair is located into one range and the number of bits to be embedded into that pixel-pair is determined by the width of this range denoted by (Wi). This width is obtained by the following equation:

$$[Wi = Ui - Li + 1]$$

while the number of bits to be embedded into that pixel-pair is calculated by [number of bits=log₂(Wi)]. Ranges close to 0 represent smooth areas and thus have smaller widths. Similarly ranges close to 255 represent clear edges and thus have larger widths. The number of bits to hide is embedded as a difference between the pixels-pair and hence the pixels values are changed accordingly.

The PVD algorithm partitions the image into blocks where each one consists of two consecutive pixels in one direction (i.e., one edge). The TPVD algorithm upgrades the capacity of the information to be hidden by partitioning the image into 2x2 blocks. Each one consists of three pixel-pair in three different directional edges (i.e., horizontal, vertical, and left diagonal). Since setting larger embedding capacity can cause image distortion, an optimal approach of selecting the reference point and branch conditions are given to achieve a minimum square error (MSE) and to reduce the effect of the stego-image distortion.

A. Embedding Phase

The embedding phase of the TPVD algorithm involves the following steps:

1. Partition the gray level of the cover image into a sequence of 2x2 blocks of pixel-pair. Fig. 1 shows one block of pixels-pair.

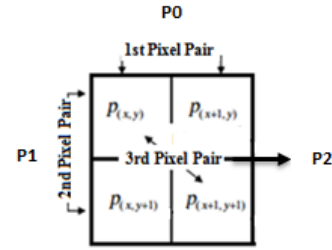


Fig.1. A 2 by 2 block of a cover image.

2. Calculate the difference for the following pixels pairs:

$$\begin{aligned} d_0 &= P_{(x+1,y)} - P_{(x,y)} \\ d_1 &= P_{(x,y+1)} - P_{(x,y)} \\ d_2 &= P_{(x+1,y+1)} - P_{(x,y)} \\ d_3 &= P_{(x,y+1)} - P_{(x+1,y+1)} \end{aligned} \quad (1)$$

3. Locate the range [Li,Ui] in the selected region for each |d_i|.
4. Compute then the width for each region as follows:

$$Wi = Ui - Li + 1 \quad (2)$$

5. Compute the amount of bits (t_i) that can be embedded in each pixel-pair as follows:

$$t_i = \log_2(Wi) \quad (3)$$

6. Check whether every t_i for the pixel-pair P₀, P₁, P₂ satisfies at least one of the following branch conditions:

$$\begin{aligned} t(P_0) \geq 5 \text{ and } t(P_1) \geq 4 \\ t(P_0) < 5 \text{ and } t(P_2) \geq 6 \end{aligned} \quad (4)$$

These conditions are used to determine the maximum amount of bits that can be embedded into every pixel-pair without deteriorating the quality of the stego-image.

- If t_i of P_i is satisfying the branch conditions, the three pixel pairs (P₀, P₁, P₂) cannot be used for embedding information. The PVD algorithm should then be used and the pixel pairs (P₀, P₄) are selected for embedding as shown in figure 2.

Fig.2. Two Consecutive pixels blocks of PVD.

- If t_i of P_i is not satisfying the branch conditions, the three pixel pairs (P_0, P_1, P_2) will be used for embedding.
7. Get the amount (t) bits from the information file and convert t to a decimal number (b).
 8. Calculate the new difference for each pixel-pair as follows :

$$\begin{aligned} d'_i &= L_i + b_i, \text{ if } d \geq 0 \\ d'_i &= -1 * (L_i + b_i), \text{ if } d < 0 \end{aligned} \quad (5)$$
 9. Modify the pixels values as follows:

$$(P'_n, P'_{n+1}) = (P_n - [m/2], P_{n+1} + [m/2]) \quad (6)$$

where $m = d'_i - d_i$
 10. If the three pixel pairs (P_0, P_1, P_2) are used for embedding, we will end up with having three different values for the pixel $p_{(x,y)}$ since it is a common pixel among the pixels-pair. Therefore, we choose the optimal reference point $p_{(x,y)}$ with the minimum MSE, and then we offset the other two pixels-pair.
 11. If the two pixel-pairs (P_0, P_4) are used for embedding the information, we calculate then t'_i of P'_0, P'_1, P'_2 . We check if t'_i is still satisfying the branch conditions. If not, we offset the values of P'_4 to satisfy these conditions.
 12. Construct a new 2×2 block and repeat the previous steps until all the information bits are embedded.

B. Extracting Phase

1. Partition the gray-scale image into 2×2 blocks of pixel pairs.
2. Calculate the difference for the pixels pairs as shown in (1).
3. Locate the range $[L_i, U_i]$ in the designed region table for each $|di|$.
4. Compute the width for each range as in (2).
5. Compute the amount of bits (t) that can be embedded in each pixel pair as in (3).
6. Check whether the computed amount of bits (t) for P_0, P_1, P_2 satisfies at least one of the branch conditions of (4). If it is satisfying then the $P_0, P_1,$

P_2 pixel pairs are selected. Otherwise, the two independent pixel pairs P_0, P_4 are selected.

7. Subtract the lower bound L_i from the $|di|$ to obtain b . Then convert b into a binary sequence with t_i bits.

III. THE PROPOSED ALGORITHM

The most challenging part of the steganography algorithms lies within choosing the appropriate pixels to embed the information. In our random-based algorithm that we denote by RTPVD we use a random factor to choose the proper combination of pixels to be used in the embedding phase. Four different combinations are used. Every combination starts with a different pixel based on a random sequence. This randomization limits the allocation of embedded information in the same direction of pixel pairs. This preserve the setgo-image quality. In addition, the information are encrypted before embedding using the well-known AES algorithm which uses 128 bits as a key and provides a high level of security. In addition, since we have a stego-key (i.e., the random generator seed) which is used to generate a random sequence for the starting pixel in every pixel-pair, it will be relatively easy to handle two keys as one (i.e., the stego key and the encryption-decryption key).

A. Block Selection

The embedding phase in the TPVD algorithm is performed using always the first pixel $P_{(x,y)}$ as a starting point. The most important thing is to have a combination with a common point included in each pixel pair to embed the information within the same direction. In the TPVD algorithm we have three possible pixel pairs where $P_{(x,y)}$ is included in each one of them. We can start by randomizing the pixel pairs into two different combinations as shown in figure 3. However, we found that the combination (i.e., figure 3a right side) is not useful to use and should then be discarded. In fact, the tri-way direction is based on the starting point and its associated pairs. To overcome this problem we choose four combinations as

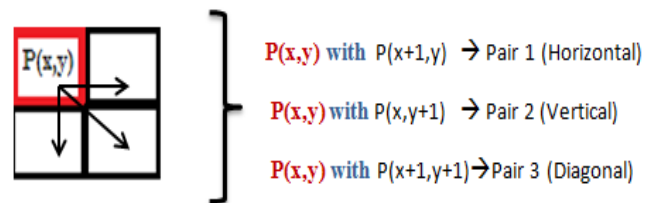


Fig.2. Pixels-pair combinations of the TPVD algorithm.

show n in figure 4. We have

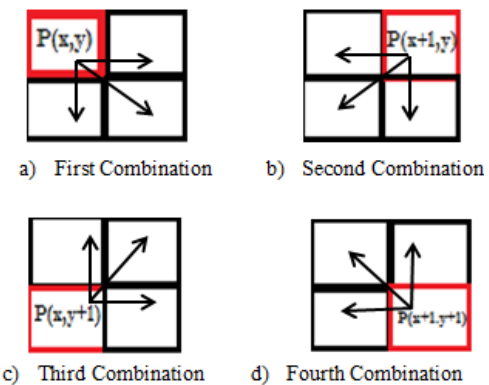


Fig. 3. All possible pixel pair combinations.

taken into consideration that each combination should have its own starting point linked to its pairs and in all possible directions (i.e., vertical, horizontal and diagonal).

B. Embedding Phase

To embed the information, we apply the same steps used in TPVD algorithm except that we randomize between different pixel pair combinations.

1. The algorithm takes as input a digital key (i.e., an integer number or seed). This key is used to generate a sequence of pixel numbers that the algorithm uses in selecting the starting pixels of every block.
2. The information are encrypted using the AES algorithm. The same key can also be used in the encryption and the decryption phase.
3. Based on the digital key, one of the proposed four pixel pair combinations is used to embed the information.
4. The key is embedded into the first block of the stego-image. It is used in the information extraction phase.

C. Extracting Phase

1. The algorithm starts by extracting the randomisation key from the first block of stego-image. Note that in this block no secret information are hidden.
2. Use the key to generate the random list of pixel pairs combinations that have been used in the embedding phase.
3. To extract the information from the stego-image, we use the same steps of the TPVD extraction algorithm except that the extraction will be done by using the random pixel pairs combinations generated.
4. The hidden encrypted information is extracted and decrypted using the AES algorithm.

IV. EXPERIMENTAL RESULTS

We have conducted two different experiments using three gray-scale images (i.e., Lena.bmp, Barbara.bmp, Girl.bmp)



Figure 5: Lena.bmp
Fig.4: Lena.bmp



Figure 6 : Barbara.bmp
Fig.5: Barora.bmp



Figure 6 : Girl.bmp
Fig.6: Girl.bmp

with 512x512 resolution. The experiments are performed with different file sizes. The text represents the information that should be hidden into the cover image.

A. First Experiment

The first experiment is performed to compare the performances of the proposed algorithm RTPVD with the TPVD algorithm. The results are presented in figures 7-10.

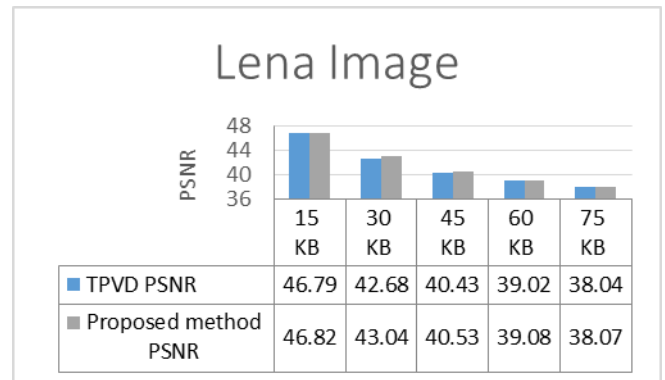


Fig. 7. First experiment for the Lena image.

The charts of the figures 7-12 represent the PSNR value with different file sizes using the TPVD and the RTPVD algorithms. Results show the PSNR values in the RTPVD algorithm are slightly better than those obtained by the TPVD algorithm for the given images of relatively small size.

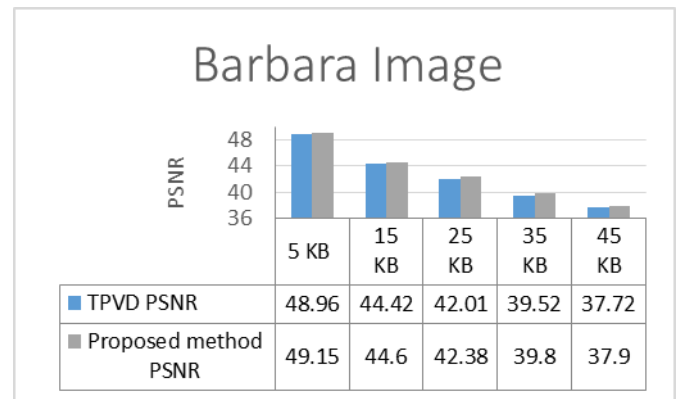


Fig. 8. First experiment for Barbara image.

B. Second Experiment

The second experiment consists of encrypting the information and then embed them into the cover image. This experiment is conducted to check whether the quality of the stego-image is preserved and to add another layer of security.

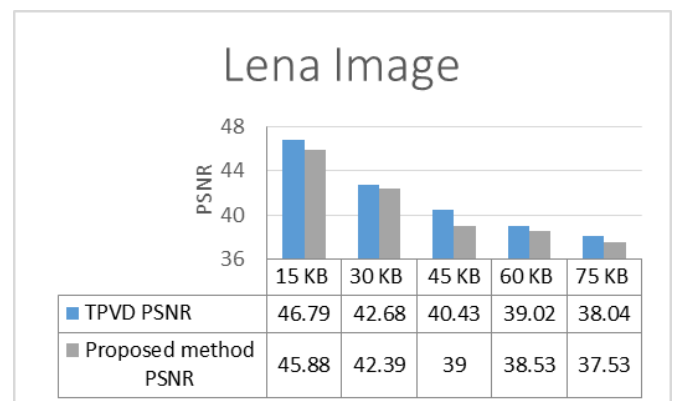


Fig.9: Second experiment for lena image.

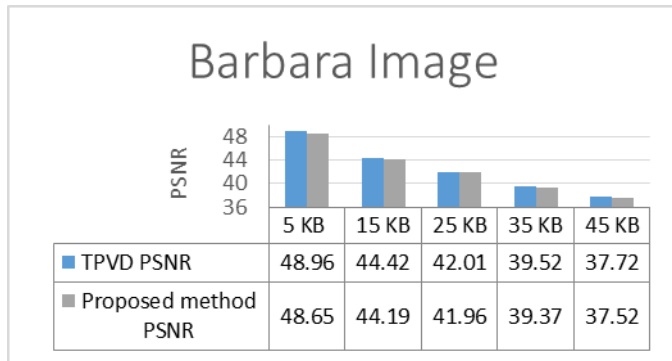


Fig. 10. Second experiment for the Barbara image.

Results show that, the PSNR value of the RTPVD is less than the PSNR of the TPVD algorithm. This is due to the fact that the AES encryption algorithm is using the block cipher technique where the message is divided into blocks. Therefore, whenever the message is smaller than the block capacity, the AES algorithm will pad the message (i.e., add bits to the message) to fit into block size. The padding process affects slightly the quality of the image.

V. CONCLUSION

We have demonstrated that the TPVD steganography image-based algorithm can slightly be improved with a random selection of the starting pixels for every pixel-pair combination. This improvement can be significant with cover images of larger sizes and may increase the amount of secret information to be hidden. We showed also that the encryption of information will result in making the algorithm more secure but can lead to the deterioration of the stego-image quality. As future work, we plan to investigate on how to determine the maximum amount of encrypted information that can be embedded into a cover image of a given size without affecting its quality.

REFERENCES

- [1] K.C. Chang, C.P. Chang, P.S. Huang, and T.M.Tu, "A novel image steganographic method using tri-way pixel-value differencing," *Journal of Multimedia*, vol. 3, no. 2, pp. 37-44, 2008.
- [2] N. Johnson and S. Jajodia, "Exploring steganography: seeing the unseen," *IEEE Computer*, pp. 26-24, 1998.
- [3] T. Morkel, "Image steganography applications for secure communication," Master Thesis, University of Pretoria, 2012.
- [4] K.G. Avinash, and M.S. Joshi, "An image steganography method with five pixel pair differencing and modulus function," *International Journal of Computer Applications*, vol. 68, no.1, pp. 20-26, 2013.
- [5] W. Hong and T. Chen, "A novel data embedding method using adaptive pixel pair matching," *IEEE Inf. Forensics Security*, vol. 7, no. 1, pp. 176-184, 2012.
- [6] A. Westfeld, "F5-A Steganographic algorithm: high capacity despite better steganalysis," In *Proceedings of the Fourth International Workshop on Information Hiding*, pp. 289-302, 2001.
- [7] D.K. Sarmah and N. Bajpai, "Proposed system for data hiding using cryptography and steganography," *Int. Journal of Computer Application*, vol. 8, no. 9, pp. 7-10, 2010.
- [8] R. Rivest, A. Shamir and L. Adleman, "A method for obtaining digital signatures and public-key cryptosystems," *Communication of the ACM*, vol. 21, no. 2, pp. 120-126, 1978.
- [9] J. Daemen and V. Rijmen, "The design of Rijndael: AES The advanced encryption standard," ISBN 3-540-42580-2 Springer-Verlag, New York, 2002.
- [10] N.I. Wu and M.S. Hwang, "Data hiding: current status and key issues," *International Journal of Network Security*, vol. 4, no. 1, pp. 1-9, 2007.
- [11] J. Korhonen and J. You, "Peak signal-to-noise ratio revisited: is simple beautiful?," *Fourth International Workshop on Quality of Multimedia Experience*, pp. 37-38, 2012.
- [12] H. Zhang, G. Geng, and C. Xiong, "Image steganography using pixel-value differencing," *Second International Symposium on Electronic Commerce and Security*, 2009.
- [13] R. Ahirwal, D. Ahirwal, and Y. K. Jain, "A high capacitive and confidentiality based Image steganography using private stego-key," in *Proceedings of the International Conference on Information Science and Applications (ISBN 978-81-907677-9-8)*, pp. 1-5, 2010.
- [14] H.W. Tseng and H.S. Leng, "A steganographic method based on pixel-value differencing and the perfect square number," *Journal of Applied Mathematics*, vol. 2013, pp. 1-8, 2013.
- [15] W. Luo, F. Huang and J. Huang, "A more secure steganography based on adaptive pixel-value differencing scheme," *Multimedia Tools and Applications*, vol. 52, issue 2-3, pp 407-430.
- [16] D.C. Wu DC, T. WH, "A steganographic method for images by pixel-value differencing," *Pattern Recognition Letter*, vol. 24, pp. 1613-1626, 2003.
- [17] X. Zhang and S. Wang, "Vulnerability of pixel-value differencing steganography to histogram analysis and modification for enhanced security," *Pattern Recognition Letter*, vol. 25, pp. 331-339, 2004.

Secure Data Sharing Polices and Architecture Preserving Privacy

Sanaa Sarahneh

Deanship of Graduate Studies and Scientific Research
Palestine Polytechnic University
Hebron, Palestine
sanaa.s@ppu.edu

Radwan Tahboub

Deanship of Graduate Studies and Scientific Research
Palestine Polytechnic University
Hebron, Palestine
radwant@ppu.edu

Abstract— Electronic data interchange can be classified as one of the important areas of information technology, where the need for data sharing increasingly required in almost every field. Data sharing concept can be defined as the process of interchanging, analyzing, retrieving and integrating data among multiple data sources in a controlled access manner. The use of information technology in different areas began to increase since 1980s; the exchange and sharing different types of information was required at that time. Although data sharing facilitates the way that data can be exchanged, security concerns arise as a challenge for conducting data sharing, many polices include confidentiality and privacy must be taken into consideration. This study will provide a literature review of security policies, focusing on privacy models to facilitate data sharing among different organizations in different areas. As a result for the study there are different data sharing model that applies different polices to preserve privacy such as Semantic Privacy-Preserving Model, Capability-based Access Control Model, and OneSwarm data sharing Model.

Keywords—*Data Sharing; Privacy; Security; Access Control; Management; Policies*

I. INTRODUCTION

Nowadays, most organizations expanded their work in the form of extranet to facilitate exchanging data among each other. Electronic data interchange can be classified as one of the important areas of information technology, where the need for data sharing increasingly required in almost every field. Data sharing can be defined as the process of interchanging, analyzing, retrieving and integrating data among multiple data sources in a controlled access manner. The use of information technology in different areas began to increase since 1980s; the exchange and sharing different types of information was required at that time. Although data sharing facilitates the way that data can be exchanged, security concerns arise as a challenge for conducting data sharing.

The remainder of this paper is organized as follows: the next section provides background in data sharing and security in data sharing. Section 3 explains three models that preserving privacy in data sharing. Next, section 4 compares between the models in terms of their advantages and disadvantages. Finally, we conclude in Section 5.

II. BACKGROUND

This section provides data sharing concepts, the need for data sharing, data sharing management, and the security for data sharing.

A. Data Sharing Concept

Data sharing concept emerges to introduce a new era of cloud computing processes, e-commerce, e-government, e-operations, e-everything. This term was coined since 1970s as [1] indicate. Reference [1] add from the early 1980s, the use of IT in the construction industry and broader engineering sector began to increase and find application in many different areas, the exchange of many different types of information was required at that time.

Reference [11] also describes data sharing as a fundamental enabler of coordination among supply chain partners. Therefore, data sharing can be defined as the process of interchanging, analyzing, retrieving and integrating data among multiple data sources in a controlled access manner, also [6] define data sharing as a fundamental to computer-supported cooperative work; people share information through explicit

communication channels and through their coordinated use of shared database.

B. The Need for Data Sharing

Reference [11] find out that data sharing is an important feature for modern organizations due to the increase in the use of communication networks, changes in architectures of enterprise information systems, as well as the increasing availability of data in computerized form, and perhaps the biggest impact on data sharing can be attributed to the widespread use of the Internet and Internet-related technologies for e-government, e-commerce, scientific research and healthcare. They add, e-government involves sharing data for transactions with citizens, other agencies and outside vendors and businesses.

Reference [11] adds, in e-commerce, data can be shared for transactions, operations, and analysis. Conducting business transactions is a basic reason for sharing data in e-commerce it is mainly used in Electronic Data Interchange (EDI), business to-business marketplaces, as well as consumer purchases over the web. They empathize that the focus of data sharing for operational purposes leads to the optimization of business processes over the entire chain to benefit all participants in the chain. Information that shared among supply chain partners may include inventory sales, demand forecasts, order status, and production schedules. Analysis, business intelligence, and decision-support represents the third purpose for data sharing in e-commerce, information available for analysis is increased through the sharing of data, they provides an example of banks data sharing with affiliates and telemarketers, another example about retailers who allow suppliers to access their inventory data for analysis purposes.

Where [10] indicates that it is highly desirable to share data among the members of the medical community; because data is very valuable, hard to produce, and in some cases irreproducible resource. Data sharing reduces the cost of reproducing redundant data collections as much as minimizing the efforts paid in performing this.

C. Data Sharing Management

Since data sharing coined, emerging data from heterogeneous sources into a single common to make data compatible with each other becomes critical issue as [7] indicate. Data integration has been attempted for about 20 years, [8] define data integration as the problem of combining the data from autonomous and heterogeneous sources, and providing users with a unified view of these data through.

Reference [13] add that many organizations and enterprises establish distributed working environment, where different users need to exchange information based on a common model, XML (eXtensible Markup Language) is used to facilitate this information exchange. The extensibility of XML allows the creation of generic models that integrate data from different sources and XML is becoming the standard format for data exchange among distributed applications components. The use of XML for information interchange among different

enterprises and organizations evokes the need for common schema that the information must follow.

D. Secure Data Sharing

Although data sharing facilitates the way that data can be exchanged, security concerns arises a challenge for conducting data sharing, confidentiality and privacy must be taken into consideration, this means, a controlled access is required to authorize authenticated users or roles to access data. Each data source represents a database, each database may use an application -for example- to access another database, this application is assigned specific permissions to access specific view of a specific database, permissions that identifies what kind of access must be granted to this application, (e.g. to read, or write, or even to have full access), for this purpose, a database of databases is needed to allow the sharing of data among the different databases as [10] indicate. Reference [2], say that this increase the need for data sharing management and data integration, on another hand data sharing and integration are prevented from being widespread because of privacy concerns, for example in e-commerce areas companies need to exchange information to boost productivity, but are prevented by fear from competitors, also sharing data in healthcare areas improve scientific research and enables early detection of disease, but without preserving privacy it is costly and difficult to make healthcare information globally expand. Reference [9] defines privacy as the process to protect information from unauthorized access.

Reference [7] say that cyber crime as well as threats to national security is costing organizations billions of dollars each year, it is equally certain that unrestricted data sharing will reduce the privacy and/or confidentiality of individuals, [7] add the challenge is to enforce appropriate administration and security policies that facilitate data sharing as needed .These policies include policies for confidentiality, privacy, and trust. During normal operations, it is important to maintain confidentiality and privacy. In addition, trust policies ensure that data is shared between trusted individuals. The standards efforts in this area include Role-based access control (RBAC) as well as Platform for Privacy Preferences (P3P) [3], also add that Public Key Infrastructure (PKI): preventing illegal modification, edits, or transfers of sensitive data to a third parties for unintended purposes.

On another hand [4] examines challenges in privacy-preserving data quality assessment, the is in protecting data and query privacy while enabling assessment of data quality held by untrusted parties. They design the protocols so that they operate on reduced dimensionality descriptions and provide a series of efficient protocols that evaluate data quality while keeping the data, the query parameters and resulting quality value private.

Another study [12] proposes a content sharing scheme that is safe in the cloud computing environment; depend on a conditional proxy re-encryption scheme. It is based on re-encryption process and the number of re-encryption keys to be required for sharing is minimized. A client is only involved in

process of encryption and decryption of data and creation of re-encryption keys.

Different models have been introduced to apply privacy in data sharing and data integration, each may be the same or different structure of other, the next section provide a literature review for preserving privacy models in different areas.

III. PRESERVING PRIVACY MODELS IN DATA SHARING

This section provides a literature review for preserving privacy models for data sharing in different areas, and their Strengths and weaknesses.

A. Semantic Privacy-Preserving Model

A semantic privacy-preserving [8] model provides authorized view-based query answering over a widespread multiple servers for data sharing and integration. For that reason model consider a large number of servers. Therefore a unified global data sharing and protection service can be achieved at the virtual platform (VP).

1) The combined semantics-enabled privacy protection policies are used to empower the data integration and access control services at the (VP). Privacy protection policies represent a long-term promise made by an enterprise to its users and is determined by business practice and legal concerns, which is expressed as combination ontology and rule:

- A privacy protection policy is a type of formal policy (FP) used for specifying a data usage constraint from a data owner. FP is a declarative expression corresponding to a human legal norm that can be executed in a computer system without causing any semantic ambiguity.
- An FP is created from a policy language (PL), and this PL is shown as a combination of ontology language and rule language.
- A formal protection policy (FPP) is an FP that aims at representing and enforcing resource protection principles, where the structure of resources is modeled as ontology's O but the resources protection is shown as rules R.(It is combination of ontology's and rules O+R).
- Semantic Web Rule Language (SWRL) Tab development tools and Semantic Query-Enhanced Web Rule Language (SQWRL), Web Ontology Language (OWL-DL) query language to model and enforce semantic privacy protection policies.

2) Three approaches have been proposed to model a set of source descriptions that specify the semantic mapping between the source schema and the global schema:

- Global-as-view (GAV) requires that the each concept in the global schema is expressed in terms of query over the data sources.

- Local-as-view (LAV) requires the global schema to be specified independently from the sources, and the source descriptions between the stable global schemas.
- Global-local-as-view (GLAV), a source description that combines the expressive power of both GAV and LAV, allowing flexible schema definitions independent of the particular details of the data sources.

3) This model is proposed with three layers, where the bottom layer provides data sources from the relational databases .The middle layer provides a semantics- enabled local schema for each independent service domain. The top layer is served at the VP, which provides a unified global view of privacy-preserving data sharing and integration services.

4) The ontology mapping and merging algorithm with a local-as-view (LAV) source description that creates a global ontology schema (mediated), which is a reconciled view of the information that provides query services to end users ,at the VP by integrating multiple local ontology schemas for data sharing. Model merged global ontology schema that mentioned above in the middle layer.

5) Using description logic (DL) to model the local and global schemas is to empower the ontology's abstract Concept representation and reasoning capabilities.

6) A query is defined as an SQWRL data log rule in the SWRL-based policy to access to a global ontology, and each SQWRL data service query for a global ontology at the VP is mapped to multiple queries as SQWRL data log rules for each local schema.

7) The challenge of designing a semantic privacy protection model is to ensure soundness and a completeness of data sharing and protection in multiple servers:

- For the soundness criterion, this model does not allow unintended data being released to the data users through the global policy schema (GPS) at the VP.
- As for the completeness criterion, the model does not miss any eligible shared data when a user asks for a data request service at the VP. Therefore, shareable data obtained the VP should equal data obtained directly from each server.

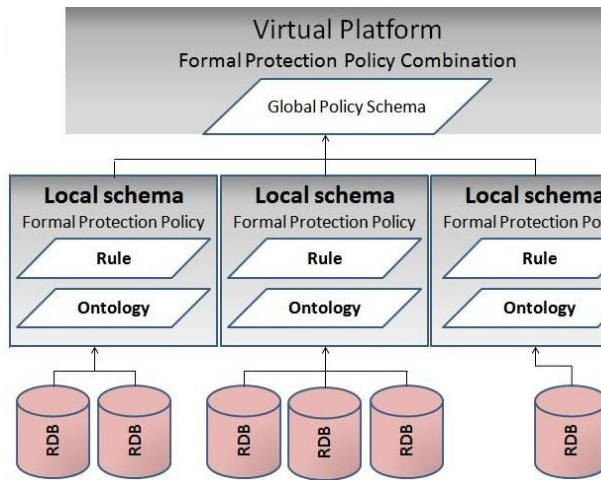


Figure (1). A semantic privacy protection model. Source: [8].

Figure (1) is proposed with three layers, where the bottom layer provides data sources from the relational databases (RDB), the middle layer provides a semantics-enabled local schema for each independent service domain. The top layer is served at the VP, which provides a unified global view of privacy-preserving data sharing and integration services. In the top layer at the VP, we have a global policy schema (GPS), including a global ontology schema (GS) aligned and merged from several local schemas (LS), e.g. TBox and a set of rule integration at the middle layer. The VP provides conceptual data access and protection services that give users a unified conceptual-global view" with access control power for each data request. Ontology-based data sources are external, independent, and heterogeneous, and each local ontology was combined with logic program (LP)-based rules for each server in the middle layer. Mapping language (ML), which semantically links a GS and integrated rule, set in the top layer to each server's ontology LS and privacy protection rules in the middle layer. Ontology and the dynamic data sources are established by defining each concept in the data sources as a view over the global schema.

• **Semantic Privacy-Preserving Strengths:**

Reference [8] list some features in semantics privacy preserving model, First ,each server shares its collected data with other servers but without breaking the original data usage commitment to its clients ,therefore a unified global data sharing and protection service can be achieved at the virtual platform (VP). Second ,the model solve the soundness and completeness of query rewriting problem using a perfect ontology merging and a perfect rule integration from the local formal protection policies, For the soundness criterion, we do not allow unintended data being released to the data users. As for the completeness criterion, we do not miss any eligible shared data when a user asks for a data request service at the VP, Third, the model develop a privacy management framework and a formal semantics language to empower agents to enforce privacy protection policies. These formal

policy using ontology for privacy protection concept descriptions and rule for data query and access control services. Ontology-based data integration in DL is to provide a uniform access mechanism to a set of heterogeneous relational database sources, freeing the user from having the knowledge about where the data are, what they are stored, and how they can be accessed.

• **Semantic Privacy-Preserving Weaknesses:**

In spite of these features, this model still have a weaknesses, it face a background policy inconsistency problem when default policy assumptions vary between different servers (one server uses open policy assumption, where no explicit option-out for data usage mean option-in, but the other server uses closed policy assumption, where no explicit option-in for data usage means option-out) and to avoid this kind of policy inconsistency by requesting all sites to use a uniform policy assumption, and to collect option-in data usage choices from users whenever multiple policies are integrated. As a conclusion Semantic Privacy-Preserving model provide secure sharing through authorized views, each organization enables data sharing and data integration without affecting its clients, but the model have inconsistency problems.

B. *Capability-based Access Control Model*

Reference [5] use a model for data sharing called Capability-based Access Control, each capability consists of a Name, which identifies a single object in the internet, and group of access rights for that object. In this model, the system sits between applications and the underlying file system. It presents applications a view-based interface to the file system. It executes queries over the local file system and communicates with other peers to evaluate distributed queries. The model is depicted through the following steps:

1) The system registers each new view and capability in a local catalog, this capability has three parts Figure (2):

- A 128-bit global view Identification ID: this ID created by concatenating a hash of the local node's Media Access Control address (MAC address) with a locally unique-for-all-time view ID, this view ID uniquely identifies an individual view in the Internet.
- A 128-bit random password: associated with each capability a 128-bit random password that ensures the capability's authenticity.
- A 32-bit IP hint field: that contains the IP address of the node that likely contains or can locate the object addressed by the capability in the Peer to Peer Network (P2P), in general, they expect that objects will not move in their network, and the IP hint will be the address of the node that created the capability and still holds its definition. If the hint fails, then it must fall back on a conventional distributed hash-table scheme for location.

128 bits	128 bits	32 bits
Global view ID	Password	IP hint

Figure (2). Capability for a view. [5].

2) The per-node catalog table generated by the system holds view and capability information. It contains two tables ViewTable and CapTable Figure (3). The ViewTable entry contains the global view ID, the view definition, and other attributes (such as the human-readable view name). For each view created on a node, there is one entry in a local view table (ViewTable). The CapTable entry stores the global view ID of the named view, the password, and the access rights. A node's capability table (CapTable) contains one entry for each capability minted to a locally known view.

Node-local view table (View Table)

Global view ID	View definition	Other attributes
...

Node-local capability table (CapTable)

Global view ID	Password	Rights
...

Figure (3). Capability and view Catalog tables. Source: [5].

3) Users grant each other access to their data simply by exchanging capabilities to their views, much like users share access to private web pages by exchanging URLs.

4) When the system receives a capability, it uses the IP hint to determine whether the capability is for a local view. If the capability is local, the system checks whether the <global view ID, password> pair in the capability matches a <global view ID, password> pair in CapTable. If so, the capability is valid, and the system then examines the access rights in CapTable to see if the requested operation is permitted. If the capability is not found in CapTable or the operation is not permitted, the request fails. If the capability is for a remote view, the system forwards the request to the appropriate node in the peer-to-peer network, which then performs the validation itself.

5) To revoke a capability, the system simply removes an entry from the CapTable. Once a capability is revoked, all queries issued on that capability will fail.

• Capability-based Access Control Strengths:

Reference [5] adds that Capability Based Access control model is a flexible protection mechanism for controlling access to shared views. Capabilities also enable rewriting and optimization of distributed queries, leading to good query execution performance. They also add, because capability is independent of the person using it, the systems access control scheme requires no user identities. Thus, sharing in a capability-based model requires no user accounts, no user authentication, and no centralized protection structure.

Capabilities facilitate data sharing because it can easily pass from user to user as a way to grant access.

• Capability-based Access Control weaknesses:

After revoking a capability, all queries issued on that capability will fail. But if a user with a capability has made a local copy of the shared data, revoking the capability cannot prevent him from distributing that copy. However, it prevents the holder from executing a query and seeing new or modified files that would result from that query. As a conclusion, the capability-based access control model provides flexible protection mechanism for controlling access to shared views, reuse of queries, it is independent of the user and decentralized.

C. Privacy-Preserving P2P Data Sharing with OneSwarm

OneSwarm [9] is a new P2P (Peer to Peer) design for data sharing that overcomes the lack of privacy in P2P data sharing applications such as BitTorrent- BitTorrent is an application that provides good performance but poor privacy- and to overcome poor performance in anonymizing overlays such as Tor. OneSwarm made a tradeoff between privacy and performance; it provides better privacy than BitTorrent and better performance than Tor.

OneSwarm builds trusted links through social network peers, instead of relying only on a directory service such as a "Tracker" that gives information to the peers about the file. OneSwarm users are free to control the tradeoff between performance and privacy by managing the level of trust.

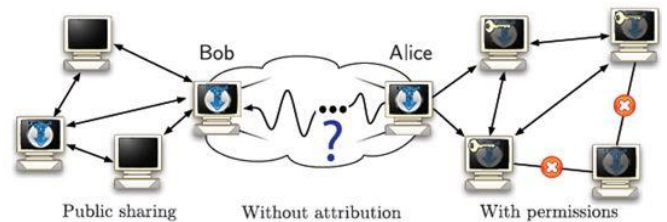


Figure (4). Cases for data sharing by OneSwarm. Source: [9].

There are three cases for OneSwarm described by [9] and shown in Figure (4), the first one is public distributed data in this case the data is not private, and direct transfers between a large set of replicas yield. The second is sharing data with permissions limits access. The last one, data shared without attribution is accessible by everyone. In public distribution anyone in the network can download file free, all data is not private, and serves as fully backwards compatible BitTorrent client. With permission case only users with permission can download files, uses persistent identities to define per file permission, this case allows all acceptable users to recognize one another. While without attribution case is depend on obscuring attribution of source and destination, it uses privacy preserving keyword search, data is relayed through unknown number of intermediaries, and it is for sensitive material.

The topology for OneSwarm the users define the links by exchanging public keys, this identifies each user and creates direct encrypted P2P connections, also OneSwarm uses social graph and community server for key distribution, Distributed Hash Table (DHT) serves as name resolution service, each client maintains encrypted entities advertising their IP address and port to authorized users, the topology is used for each transfer. In each transfer each OneSwarm client restricts direct communication to a small number of persistent contacts and locates different data sources using object lookup through overlay, this topology is used to enhance privacy, while to enhance performance in OneSwarm protocol, multiple paths to each data source are used.

Linking peers with trust relationships is explained by [9] it uses 1024 bit RSA (Rivest-Shamir-Adleman cryptosystem) public/private key pair which is generated in installation phase, public key serves as its identity among friends, manual key sharing between two users; the automatic key sharing discovers and exchange keys over local area network or by email invitation to friends. Managing untrusted peers by private community server and public community server, the private is to maintain a list of registered users and to provide authorized subscribers with a current set of public keys, the public is to allow new users to easily obtain a set of untrusted peers. Identity in OneSwarm protocol are managed by the DHT which contain of hashed IP and port, entries for a client encrypted with the public key, each entry is indexed by 20 byte randomly generated.

Naming and locating data in OneSwarm used Secure Sockets (SSLv3) for connection as [9] say, file list messages is exchanged on first connection then compressed XML attributes which contain name, size and other meta data for particular peer. Shared files are named using 160 bit SHA-1 hash, for public data user obtains hashes from email, websites and keywords search, while for private data user must obtain both hash and key used for decryption of data. The risk in OneSwarm model as [9] describe, the aattacker can join with limited number of nodes, also can check the traffic flow to/from, also may sniffing, modify or injected data. Limiting hacker to snoop in from by not assigning peer dynamically, also defining trusted and untrusted links to keep the information private, end to end path between users change rapidly helps to prevent hacking using historical data. [9] adds, preventing timing attack by search queries and responses are forwarded after adding a random delay to inhibit calculation of round trip time (RTT) to infer proximity, preventing correlation attack by having limited view of the overlay and cannot control path setup beyond directly connected neighbors, attackers could use this to correlate performance with ongoing transfers, finally preventing collusion attack by search queries and responses are forwarded probabilistically, making it very hard for directly connected colluding peers to infer source of data or monitor habits.

- **OneSwarm Data Sharing Model Strengths:**

OneSwarm provides flexibility for the user to manage the level of privacy for file sharing, incorporation of social network for building P2P file sharing network, and reduce cost of privacy.

- **OneSwarm Data Sharing Model Weaknesses:**

There are Delayed responses to queries from untrusted peers.

IV. COMPARISON BETWEEN THE MODELS

The following table provides a comparison between the privacy preserving models in terms of their advantages and disadvantages:

Based on the comparison between the three models, the Capability-based Access Control model has disadvantages, mainly: there is no fixed method for translation, difficulties in integrity control, cannot prevent the user from keeping and distributing the shared data, and decentralized control. These disadvantages make the implementation of the model hard, concerning semantic privacy preserving model overcomes the previous disadvantages, and provides data integration, secure sharing through authorized view, in addition, each organization enable data sharing without affecting its clients, while OneSwarm data sharing model provides flexibility for the user to manage the level of privacy for file sharing, and reduces cost of privacy but it has delay in response.

V. CONCLUSION

Data sharing concept can be defined as the process of interchanging, analyzing, retrieving and integrating data among multiple data sources in a controlled access manner. Although data sharing facilitates the way that data can be exchanged, security concerns arises a challenge for conducting data sharing, many polices include confidentiality and privacy must be taken into consideration. In this study we provide a literature review of security policies, focusing on privacy models that facilitate data sharing among different organizations in different areas. As a result for the study there are different data sharing model that applies different polices to preserve privacy, and semantic privacy preserving model overcomes many disadvantages of others models, and provide data integration, secure sharing through authorized view, in addition, each organization enable data sharing without affecting its clients.

TABLE (1). MODELS COMPARISON

Model Name	Advantages	Disadvantages
Semantic Privacy-Preserving model	1. Each organization enables data sharing without affecting its clients. 2. Data integration. 3. Provide secure sharing through authorized views.	1. Inconsistency problems.
Capability-based Access Control	1. Provide flexible protection mechanism	1. Cannot prevent the user from

model	for controlling access to shared views. 2. Reuse of queries. 3. Independent of the user.	keeping and distributing the shared data. 2. Decentralized control.
OneSwarm Model	1. Efficient, robust. 2. Users flexible.	1. There are delayed responses to queries from untrusted peers.

REFERENCES

[1] Bakis, N., Aouad, G., Kagioglou, M., (2007), "Towards distributed product data sharing environments Progress so far and future challenges", *Elsevier-Automation in Construction*, 16, (5): 586-595.

[2] Clifton, C., Doan, A., Elmagarmid, A., (2004), "Privacy Preserving Data Integration and Sharing", *ACM- Research issues in data mining and knowledge discovery*: 19-26.

[3] Choi, J., Chun, S., Kim, D., Keromytis, A., (2013), "SecureGov: Secure Data Sharing for Government Services", *The Proceedings of the 14th Annual International Conference on Digital Government Research*.

[4] Freudiger, J., Rane, S., Brito, A., Uzun, E., (2014), "Privacy Preserving Data Quality Assessment for High-Fidelity Data Sharing", *Proceedings of the 2014 ACM Workshop on Information Sharing & Collaborative Security*, 21-29.

[5] Geambasu, R., Balazinska, M., Gribble, S., Levy, H., (2007), "HomeViews: Peer-to-Peer Middleware for Personal Data Sharing Applications", *ACM*: 235-246.

[6] Greif, I., Sarin, S., (1987), "Data Sharing in Group Work", *ACM*, 5, (2): 187-211.

[7] Harris, D., Khan, L., Paul, R., Thuraisingham, B., (2007), "Standards for secure data sharing across organizations", *ACM-Computer Standards and Interfaces*, 29,(1): 86-96.

[8] Hu, Y., Yang, J., (2011), "A Semantic Privacy-Preserving Model for Data Sharing and Integration", *ACM-Web Intelligence, Mining and Semantics*, (9): 1-12.

[9] Isdal, T., Piatek, M., Krishnamurthy, A., Anderson, T., (2010), "Privacy-Preserving P2P Data Sharing with OneSwarm", *ACM SIGCOMM Computer Communication Review*, 40, (4): 111-122.

[10] Mannai, D., Bugarara, K., (1993), "Enhancing Inter-Operability and Data Sharing In Medical Information Systems", *ACM*, 22, (2): 495-498.

[11] Sarathy, R., Muralidhar, K., (2004), "Secure and useful data sharing", *Elsevier*, 42, (1): 204- 220.

[12] Son, J., Kim, H., Kim, D., (2014), "On Secure Data Sharing in Cloud Environment", *Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication*, 6.

[13] Varlamis, I., Vazirgiannis, M., (2001), "Bridging XML-Schema and relational databases A system for generating and manipulating relational databases using valid XML documents", *ACM*,: 105 - 114.

Groebner Bases and Coding

Thomas Risse

Institute of Informatics and Automation
 Hochschule Bremen, University of Applied Sciences
 Bremen, Germany
risse@hs-bremen.de

Abstract— In the past Groebner bases have been proved to be a very potent tool to solve a variety of problems first of all in mathematics but also in science and engineering. Hence, it is near at hand to study application of Groebner bases in coding, i.e. the encoding and especially the decoding of linear error correcting codes. This paper attempts an overview focusing on Reed-Solomon codes and Goppa codes together with their coding and decoding algorithms.

Keywords— Groebner bases, error correcting codes, decoding, Reed-Solomon codes, Goppa codes

I. INTRODUCTION

A Groebner basis (according to Bruno Buchberger, 1965) or a standard basis (according to Heisuke Hironaka, 1964) is a finite generating set of an ideal I in the polynomial ring $R = K[x_1, \dots, x_n]$ over a field K . For any such ideal the (reduced) Groebner basis is unique and can be determined algorithmically. This basis allows solving some prominent mathematical problems, e.g. to decide whether some polynomial belongs to I or not, whether two ideals are identical, whether two varieties are identical or not etc. In his seminal thesis [5] Buchberger developed the theory and presented the necessary algorithms. He also investigated applications [7] of Groebner bases like solving systems of multivariate polynomial equations.

In the eighties the rapid development of computers spurred further investigation of Groebner bases which resulted in improvements of the algorithms and even more applications [7]. Especially Computer Algebra Systems benefitted [9] from Groebner bases. But Groebner bases also brought forth progress in coding and cryptography.

II. GROEBNER BASES

A. Definitions [11]

Let I be some ideal in $R = K[x_1, \dots, x_n]$. Then by Hilberts basis theorem, I is finitely generated, i.e. $I = \langle f_1, \dots, f_s \rangle$. Now fix some monomial order on the monomials in R to be able to specify leading monomials $LM(f)$, leading terms $LT(f)$ and leading coefficients $LC(f)$ for any f in R . Then a Groebner basis G for I is a set $G = \{g_1, \dots, g_t\}$ with $I = \langle G \rangle$ so that the ideal generated by the leading terms of the elements in I is generated by the leading terms $LT(g)$ for g in G , i.e. $\langle LT(g_1), \dots, LT(g_t) \rangle = \langle LT(I) \rangle$. Equivalently, $G = \{g_1, \dots, g_t\}$ is a Groebner basis if and only if $LT(f)$ is divisible by one of the $LT(g_i)$ for any f in I . By the way, G then has the minimality properties of a proper ideal basis. Furthermore, G is

unique, and any f in R can be written as $f = g + r$ for unique polynomials g and r with g in I and no term of r is divisible by any element of $LT(g_i)$.

B. Algorithms [11]

Buchberger's algorithm computes a (not reduced) Groebner basis for an ideal $I = \langle f_1, \dots, f_s \rangle$ using syzygy- or S-polynomials $S(f, g) = \frac{LCM(LM(f), LM(g))}{LT(f)} f - \frac{LCM(LM(f), LM(g))}{LT(g)} g$ for any two polynomials f and g in R together with a generalization of the polynomial division algorithm for polynomials in one variable to the case of multivariate polynomials f, f_1, \dots, f_s, r in R such that $f = a_1 f_1 + \dots + a_s f_s + r$ where the remainder $r = \bar{f}^{(f_1, \dots, f_s)}$ is zero or a K -linear combination of monomials none of which is divisible by any $LT(f_1), \dots, LT(f_s)$ – all in the usual notation of [11][21] et al. With these definitions Buchberger's algorithm can now be specified.

```

input:  $F = (f_1, \dots, f_s) \subset K[x_1, \dots, x_n]$ 
output:  $G = (g_1, \dots, g_t)$  with  $\langle F \rangle = \langle G \rangle$ 
repeat
     $G' := G$ 
    for each  $\{p, q\} \subset G', p \neq q$  do
         $S := \overline{S(p, q)}^G$ 
        if  $S \neq 0$  then  $G := G \cup \{S\}$ 
until  $G = G'$ 
    
```

Code snippet 1. Computation of G with $\langle G \rangle = \langle F \rangle$

Obviously, this very simple version of Buchberger's algorithm extends the given set F to G . A reduction step removes superfluous elements from G resulting in the unique reduced Groebner basis of I . There are improved versions [11] to compute the unique, reduced Groebner basis of I efficiently.

C. Applications [11]

First, one should note that the concept of Groebner bases generalizes both Euclid’s algorithm to compute the greatest common divisor, gcd, of two polynomials as well as Gauß’s algorithm to solve a system of linear equations.

Euclids algorithm

```

in:  $f, g \in K[x]$ ; out:  $h = \text{gcd}(f, g)$ 
 $h := f; s := g;$ 
while  $s \neq 0$ 
     $r = \text{remainder}(h, s);$ 
     $h := s; s := r;$ 
```

Code snippet 2. $h = \text{gcd}(f, g)$ for any $f, g \in K[x]$

Regard each equation of a system of linear equations in the unknowns x_1, \dots, x_n as a linear polynomial f_i in $K[x_1, \dots, x_n]$. Then, the reduced Groebner basis $G = \{g_1, \dots, g_t\}$ of $I = \langle f_1, \dots, f_n \rangle$ consists of linear, non-zero polynomials whose coefficients correspond to the non-zero rows in the reduced echelon form of the coefficient matrix of the system of linear equations. In this sense, computation of the reduced Groebner basis is equivalent to Gauß’s algorithm.

As one of the very many applications of Groebner bases consider the problem to solve a system of multivariate polynomial equations $f_1 = f_2 = \dots = f_s = 0$ for f_i in R . Here we use $I = \langle f_1, \dots, f_s \rangle = \langle G \rangle$ for the reduced Groebner basis $G = \{g_1, \dots, g_t\}$ of I . It turns out that the set of equations $g_1 = g_2 = \dots = g_t = 0$ is easier to solve because using the lexicographic order (*lex*), variables are eliminated in that order in the Groebner basis so that a process like back substitution generates the variety $V(I) = V(\langle G \rangle) = V(g_1, \dots, g_t)$. Elimination theory [11] provides the proofs and [20] more examples.

III. ERROR CORRECTING CODES

Here we consider linear block codes [17] only. An alphabet is some finite field $\mathbb{F} = \mathbb{GF}(p^m)$ for prime p and $m \in \mathbb{N}$ and information words u of length k are in \mathbb{F}^k . Code words over \mathbb{F} are of the form uG for a $n \times k$ generator matrix G . Hence the code $\mathcal{C} = \{uG : u \in \mathbb{F}^k\}$ is a linear subspace of \mathbb{F}^n . \mathcal{C} can also be characterized as the kernel space $\mathcal{C} = \{c \in \mathbb{F}^n : Hc^T = 0\}$ of the parity matrix H , i.e. $HG^T = 0$. If any two code words have a Hamming distance of at least d then at most $(d - 1)/2$ errors in a transmitted code word can be corrected. Such a code is called a (linear) $[n, k, d]$ code.

Encoding an information word u to $c = uG \in \mathcal{C}$ is easy whereas decoding a corrupted word $y = c + e$ with an error vector $e \in \mathbb{F}^n$ with no more than $(d - 1)/2$ non-zero elements to the original c (and then to the original information word u) is difficult. In fact, it is NP-complete [2]. However, for many specific (linear) codes there exist efficient decoding algorithms.

A. (Generalized) Reed-Solomon Codes

(Generalized) Reed-Solomon codes, RS and gRS, are an important class of codes comprising many other important

codes. Such code \mathcal{C}_{gRS} is specified by its n distinct non-zero code locators $\alpha_1, \dots, \alpha_n \in \mathbb{F}$ and n column multipliers $v_1, \dots, v_n \in \mathbb{F}$. Then the parity matrix H_{gRS} of \mathcal{C}_{gRS} is defined by

$$H_{gRS} = \begin{pmatrix} \alpha_1^0 & \alpha_2^0 & \dots & \alpha_n^0 \\ \alpha_1^1 & \alpha_2^1 & \dots & \alpha_n^1 \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_1^{n-k-1} & \alpha_2^{n-k-1} & \dots & \alpha_n^{n-k-1} \end{pmatrix} \begin{pmatrix} v_1 & & & 0 \\ & v_2 & & \\ & & \ddots & \\ 0 & & & v_n \end{pmatrix}$$

Then, \mathcal{C}_{gRS} is a (linear) $[n, k, d]$ code with $d = n - k + 1$. (Such codes attain the Singleton bound $d \leq n - k + 1$ and are called maximum distance separable, MDS codes.) For gRS codes there are efficient decoding algorithms: e.g. solving linear equations [17], using Euclid’s algorithm [22] or linear recurrences in case of the famous Berlekamp-Massey algorithm [3][17][18]. List decoding of e.g. (generalized) Reed-Solomon codes relaxes the assumption on the number of allowed errors and returns a list of possible code words.

B. Goppa-Codes

Goppa-codes, alternant gRS codes, play an important role e.g. in the McEliece Public Key Crypto System, PKCS [18][19]. Let $F = \mathbb{GF}_q$, $K = \mathbb{GF}(q^m)$ and $L = \{\alpha_1, \dots, \alpha_n\} \subset K$ be a set of pair wise different code locators and let $g(x) \in K[x]$ with $0 \notin g(L)$ be a Goppa-polynomial of degree t . Then

$$\mathcal{C}_{Goppa} = \{(c_1, \dots, c_n) \in F^n : \sum_{i=1}^n \frac{c_i}{x - \alpha_i} = 0 \text{ mod } g(x)\}$$

is a linear $[n, k, d]$ code over F . The code \mathcal{C}_{Goppa} is called irreducible iff the Goppa polynomial g is irreducible. Let $g(x) = \sum_{i=0}^t g_i x^i$ be the Goppa polynomial. Then we have (best shown by induction in t , the degree of g) $\frac{g(x)-g(\alpha)}{x-\alpha} = g_t \sum_{i=0}^{t-1} \alpha^i x^{t-1-i} + g_{t-1} \sum_{i=0}^{t-2} \alpha^i x^{t-2-i} + \dots + g_2(x + \alpha) + g_1$

Then, $c \in \mathcal{C}_{Goppa}$ iff $\sum_{i=1}^n \frac{c_i}{g(\alpha_i)} \frac{g(x)-g(\alpha)}{x-\alpha} = 0$ in $K[x]$ and by comparison of coefficients $c \in \mathcal{C}_{Goppa}$ iff $Hc^T = 0$ with parity matrix

$$H = \begin{pmatrix} \frac{g_t}{g(\alpha_1)} & \frac{g_t}{g(\alpha_2)} & \dots & \frac{g_t}{g(\alpha_n)} \\ \frac{g_{t-1} + \alpha_1 g_t}{g(\alpha_1)} & \frac{g_{t-1} + \alpha_2 g_t}{g(\alpha_2)} & \dots & \frac{g_{t-1} + \alpha_n g_t}{g(\alpha_n)} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{g_1 + \alpha_1 g_2 + \dots + \alpha_1^{t-1} g_t}{g(\alpha_1)} & \frac{g_1 + \alpha_2 g_2 + \dots + \alpha_2^{t-1} g_t}{g(\alpha_2)} & \dots & \frac{g_1 + \alpha_n g_2 + \dots + \alpha_n^{t-1} g_t}{g(\alpha_n)} \end{pmatrix} =$$

CXY where

$$C = \begin{pmatrix} g_t & 0 & \dots & 0 \\ g_{t-1} & g_t & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ g_1 & g_2 & \dots & g_t \end{pmatrix}, X = \begin{pmatrix} \alpha_1^0 & \alpha_2^0 & \dots & \alpha_n^0 \\ \alpha_1^1 & \alpha_2^1 & \dots & \alpha_n^1 \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_1^{t-1} & \alpha_2^{t-1} & \dots & \alpha_n^{t-1} \end{pmatrix},$$

and $Y = \begin{pmatrix} \frac{1}{g(\alpha_1)} & & & \\ & \frac{1}{g(\alpha_2)} & & \\ & & \ddots & \\ & & & \frac{1}{g(\alpha_n)} \end{pmatrix}$.

Such codes correct up to $\frac{t}{2}$ errors, even up to t errors in the binary case, i.e. if \mathcal{C}_{Goppa} is a code over $\mathbb{F} = \mathbb{GF}(2)$.

Early methods used Euclid's algorithm for decoding or list decoding [23]. Later Patterson's algorithm [16] provided an efficient method to decode received words when using a Goppa encoding [18]. On top one can correct approximately up to $\frac{t^2}{n}$ errors [4].

C. Cyclic Codes

Cyclic codes [17] are linear codes \mathcal{C} when in addition with any code word $(c_0, \dots, c_{n-1}) \in \mathcal{C}$ also $(c_{n-1}, c_0, \dots, c_{n-2}) \in \mathcal{C}$, i.e. the shifted word is again a code word. Conventional Reed-Solomon codes (code locators $\alpha_j = \alpha^{j-1}$ are powers of an element $\alpha \in \mathbb{F}$ of multiplicative order n) as well as BCH codes (alternant codes of conventional Reed-Solomon codes) are prominent examples of cyclic codes. Cyclic codes feature efficient encoding (multiplication by the generator polynomial g of the code), syndrome computation (remainder of division by g) and decoding (sequentially by Meggitt decoder) via rather simple hardware.

IV. APPLYING GROEBNER ALGORITHMS TO CODING

There are several ways [10] to transform the decoding problem into a problem of solving a system of multivariate polynomial equations. A straightforward way is to consider the (unknown) entries e_i of the error vector e as variables E_i . If H consists of rows h_1, \dots, h_r with redundancy $r = n - k$ then the vector equation $s = He^T$ is equivalent to the r linear equations

$$\sum_{j=1}^n (h_i)_j E_j - s_i = 0 \text{ for } i = 1, \dots, r \quad (1)$$

We can formulate the condition that e has at most $t = \lfloor \frac{d-1}{2} \rfloor$ non-zero entries by the $\binom{n}{t+1}$ equations of multidegree $t + 1$

$$E_{j_1} \cdot E_{j_2} \cdot \dots \cdot E_{j_{t+1}} = 0 \text{ for } 1 \leq j_1 < j_2 < \dots < j_{t+1} \leq n \quad (2)$$

Let the two sets of equations together generate the ideal I . Then the Groebner basis of I allows to read off the solution E , $E = (E_1, \dots, E_n)$ i.e. the one element in the variety $V(I)$.

In addition, [10][21] present alternatives to (2) with less equations of lower multidegree so that the Groebner basis is faster to compute.

A. RS and gRS codes

Decoding RS and gRS codes means to solve the key equations. Hence in general a formulation of the decoding problem using Groebner bases is near at hand. But exploiting the fact that Groebner bases help to determine the corresponding variety $V(I)$ of some ideal $I = \langle G \rangle$ for the reduced Groebner basis G of I explains why Groebner bases support list decoding naturally. [15] gives an overview over existing methods.

B. Goppa codes

[14] is most promising to decode Goppa codes. However, [14] shows 'that one can, at least in theory, decode these codes up to half the true minimum distance by using the theory of Groebner bases'. Therefore, what is lacking is the transfer of the solution of [14] into praxis.

C. Cyclic codes

[12] gives an algorithm to decode cyclic codes using Groebner bases. The decoding problem is represented as a system of $n - k$ linear equations together with n quadratic equations in at most $n + d$ unknowns, i.e. error locations and error values. Because the number of errors is not known beforehand, the algorithm then starts with assumed $t = 0$ errors and increases t as long as the variety $V(I) = \emptyset$ where I is the ideal generated by equations specified above. Once $V(I) \neq \emptyset$ it contains the unique solution. However, the viability of the algorithm is limited because on one hand there are aforesaid efficient decoding methods and on the other hand the cost to compute a Groebner basis might be prohibitive.

CONCLUSION

This article is meant to set the stage for Groebner bases in coding. In the light of the very many application of Groebner bases in science and engineering [7] it is to be expected that further research will reveal even better algorithms for the decoding of linear (and non-linear) error-correcting codes. (Also, Groebner bases have spurred the specification and investigation of new linear codes [13][14].) The exact average complexity of determining the reduced Groebner basis of an ideal is not known right now. Once it has been determined [10] one will be able to set objectives and to identify limits of the approach to apply Groebner bases for coding.

REFERENCES

- [1] E.R. Berlekamp, Goppa Codes, IEEE Transactions Information Theory, Vol 19, No 5, September 1973, 590–592
http://infosec.seu.edu.cn/space/kangwei/senior_thesis/Goppa.pdf
- [2] .R. Berlekamp, R.J. McEliece, H.C.A. van Tilborg, On the inherent intractability of certain coding problems, IEEE Trans. Inform. Theory, 24 (1978) 384–386
- [3] E.R. Berlekamp, Algebraic Coding Theory, Aegean Park Press 1984
- [4] D.J. Bernstein, List decoding for binary Goppa codes, 2008
<http://cr.yp.to/ntheory/goppalist-20080706.pdf>
- [5] B. Buchberger, Ein Algorithmus zum Auffinden der Basiselemente des Restklassenringes nach einem nulldimensionalen Polynomideal, Universität Innsbruck, Dissertation, 1965
http://www.ricam.oeaw.ac.at/Groebner-Bases-Bibliography/gbbib_files/publication_706.pdf
- [6] B. Buchberger, An Algorithmic Criterion for the Solvability of a System of Algebraic Equations, Aequationes Mathematicae 4 (1970), 374–383
http://www.ricam.oeaw.ac.at/Groebner-Bases-Bibliography/gbbib_files/publication_699.pdf
- [7] B. Buchberger, Groebner Bases – A Short Introduction for System Theorists, Proceedings of Computer Aided Systems Theory, EUROCAST, 1–19, 2001
<http://people.reed.edu/~davidp/pcmi/buchberger.pdf>

- [8] B. Buchberger, H. Engl, Workshop D1: Groebner Bases in Cryptography, Coding Theory, and Algebraic Combinatorics, May 1st – May 6th, 2006 with papers on coding, cryptography, algebraic combinatorics, etc. http://www.ricam.oeaw.ac.at/specsem/srs/groeb/schedule_D1.html
- [9] B. Buchberger, A. Maletzky (organizers), session ‘Software for Groebner Bases’, 4th Int. Congress on Mathematical Software, ICMS, Seoul, August 5th – 9th, 2014
<http://www.risc.jku.at/people/amaletzky/ICMS2014-GB.html>
- [10] S. Bulygin, R. Pelikaan, Bounded distance decoding of linear error-correcting codes with Groebner bases, *J. Symbolic Computation* 44 (2009) 1626–1643
<http://www.sciencedirect.com/science/article/pii/S0747717108001776>
- [11] D. Cox, J. Little, D. O’Shea, *Ideals, Varieties, and Algorithms*, Springer 2007
<http://www.math.ku.dk/~holm/download/ideals-varieties-and-algorithms.pdf>
- [12] M. de Boer, R. Pelikaan, Gröbner bases for error-correcting codes and their decoding.
http://www.risc.jku.at/Groebner-Bases-Bibliography/gbbib_files/publication_590.pdf
- [13] C. Di, Z. Liu, Construction of a Class of Algebraic-Geometric Codes via Groebner Bases, *MM Research Preprints*, 42–48 No. 16, April 1998. Beijing
<http://www.mmrc.iss.ac.cn/pub/mm16.pdf/cdi.pdf>
- [14] J. Fitzgerald, R.F. Lax Decoding affine variety codes using Groebner bases, *Designs, Codes and Cryptography*, 13, 147–158 (1998)
<https://www.math.lsu.edu/~lax/designscodescrypt.pdf> or
http://www.ricam.oeaw.ac.at/Groebner-Bases-Bibliography/gbbib_files/publication_277.pdf
- [15] H. O’Keefe, P. Fitzpatrick, A Groebner basis approach to list decoding of Reed-Solomon and Algebraic Geometry Codes, see [8]
<https://www.ricam.oeaw.ac.at/specsem/srs/groeb/download/OKeefe.pdf>
- [16] N. J. Patterson: The Algebraic Decoding of Goppa Codes; *IEEE Trans. on Information Theory*, Vol IT-21, No 2, March 1975, 203–20
- [17] R.M. Roth, *Introduction to Coding Theory*, Cambridge 2006
- [18] Th. Risse, How SAGE helps to implement Goppa Codes and McEliece PKCSs, *Int. Conf. on Information Technologies 2011, ICIT’11*, May 11th – 13th, 2011, Amman
<http://www.weblearn.hs-bremen.de/risse/papers/ICIT11/>
- [19] Th. Risse, Generating Goppa Codes, *Int. Conf. on Information Technologies 2013, ICIT’13*, May 8th – 10th, 2013, Amman
<http://sce.zuj.edu.jo/icit13/index.php/accepted-papers/2-uncategorised/41-applied-mathematics>
- [20] Th. Risse, Groebner-Basen, 12. Workshop Mathematik für Ingenieure, Hafen City Universität, Hamburg 12. – 13.2.2015,
<http://www.weblearn.hs-bremen.de/risse/papers/MathEng12/>
- [21] M. Sala, T. Mora, L. Perret, S. Sakata, C. Traverso (Editors), *Groebner Bases, Coding, and Cryptography*, Springer 2009
<http://xa.yimg.com/kq/groups/24626876/489439549/name/ggp496.pdf>
- [22] P. Shankar, Decoding Reed–Solomon Codes Using Euclid’s Algorithm, *Resonance* April 2007, 37–51
<http://www.ias.ac.in/resonance/Volumes/12/04/0037-0051.pdf>
- [23] Y. Sugiyama, M. Kasahara, S. Hirasawa, T. Namekawa, A Method for Solving Key Equation for Decoding Goppa Codes, *Information and Control* 27 (1975) 87–99
<http://www.sciencedirect.com/science/article/pii/S001999587590090X>

Hide Image in Image Based on LSB Replacement and Arnold Transform

Dr. Sadik Ali Al-Taweel
Information Systems Department
University science and Technology
Sana'a, Yemen
Dr.sadiq@ust.edu

Mohammed. Al-Hada, Ahmed. M. A. Naser, and Mohammed Al-Thamary
Information Technology Department
University science and Technology
Sana'a, Yemen
alhada.eng@gmail.com

Abstract— For quite a long time, computer security was a rather narrow field of study that was populated mainly by theoretical computer scientists, electrical engineers, and applied mathematicians. Data hiding techniques have taken important role with the rapid growth of intensive transfer of multimedia content and secret communications. There are many techniques used for data hiding and the well-known technique is the Steganography. Steganography is the art of hiding information in ways that prevent detection. For hiding secret information in images, there exists a large variety of Steganography techniques, some are more complex than others and all of them have respective strong and weak points. In this paper deals with encrypt and hide image in another gray image file using Least Significant Bit (LSB) based Steganography and Arnold's transformation algorithm based Cryptography. Experimental results show that the algorithm has good security and imperceptibility in grayscale images.

Keywords— Image processing, Steganography, information hiding, Arnold Transform

I. INTRODUCTION

Based on [1] stated that "Security through obscurity says that if you hide the inner workings of your system you will be secure. This philosophy does not work when it comes to security, and it does not work when it comes to cryptography". Most of the requirements of secret communication, sometimes in combination with other techniques, such as cryptography, as cryptography and Steganography complement each other. It is recommended to use these two techniques together for a higher level of security.

Information security is the protection of information and the systems and hardware that use, store, and transmit that information. Information security can be defined as measures adopted to prevent the unauthorized use or modification of use of data or capabilities [2].

Image scrambling refers to some kind of transform, which makes the spatial location of the pixel becomes chaos and lost their original features and meaning. But the total number of pixels and histogram has not unchanged so as to achieve the purpose of encryption. As well as the scrambling must be one kind reversible transform, otherwise it will not have any significance in the practical application. If you do not know the rules and keys of the transform, it is impossible to recover the

original image. And in the process of scrambling, the loss of the hidden information is dispersed into the whole hidden data. Thereby it minimizes the loss of meaningful information in order to reach purpose of improving robustness. Therefore image scrambling technology has been widely applied in the image Steganography field [3].

Arnold transform is a type of image scrambling methods. The transformation shifts pixel position from (x, y) to (x', y') without changing its gray value. It is cyclic the secret image repeats itself after certain number of iteration.

II. RELATED WORK

M. Mahdavi, et.al [4] proposed a new accurate steganalysis method for the LSB replacement Steganography. The suggested method is based on the changes that occur in the histogram of an image after the embedding of data. Every pair of neighboring bins of a histogram are either inter-related or unrelated depending on whether embedding of a bit of data in the image could affect both bins or not.

Chang, C.C et al [5] proposed an image Steganography technique which offer high embedding capacity and bring less distortion to the stego image. The embedding process embed bits of secret bit stream on the stego image pixels. Instead of replacing the LSB of every pixel, this method replaces the pixel

intensity with similar value. The range of modifiable pixel value is higher in edge areas than smooth areas to maintain good perceptual excellence. This method is falling of boundary problem which means the pixel which is located for embedding will become unused; since it exceeds the maximum intensity level which is greater than 255 (maximum gray scale intensity).

Wu. H.C et al [6] suggested to improve the capacity of the hidden secret data and to provide an imperceptible Stego-image quality. This method based on least significant bit (LSB) replacement and pixel-value differencing (PVD) method is presented in this paper. The limitation of this method is Low-hiding capacity owes to mainly hiding in smooth areas. For example if pixel value difference is 3 if the corresponding range width is 8, only 3 bits can be embedded in a pair of pixels.

Y. K. Jain et al [7] proposed an adaptive least significant bit spatial domain embedding method. The proposed method divides the image pixels ranges (0-255) and generates a stego-key. This private stego-key has 5 different gray level ranges of image and each range indicates to substitute fixed number of bits to embed in least significant bits of image. The strength of this method is its integrity of secret hidden information in stego-image and high hidden capacity. This method could be weak for the hide extra bits of signature with hidden message for its integrity purpose. This study also proposed a method for color image just to modify the blue channel with this scheme for information hiding. This method is targeted to achieve high hidden capacity plus security of hidden message.

According to Yang et al., in [8], an adaptive LSB substitution based data hiding method for image is proposed, to achieve better visual quality of stego-image. It takes care of noise sensitive area for embedding. Proposed method differentiates and takes advantage of normal texture and edges area for embedding. This method analyzes the edges, brightness and texture masking of the cover image to calculate the number of k-bit LSB for secret data embedding. The value of k is high at non-sensitive image region and over sensitive image area (k) value remain small to balance overall visual quality of image. The LSB's (k) for embedding is computed by the high-order bits of the image. It also utilizes the pixel adjustment method for better stego-image visual quality through LSB substitution method. The overall result shows a good high hidden capacity, but dataset for experimental results are limited, there is not a single image which has many edges with noise region like 'Baboon.tif'.

C.-H. Yang et al [9] proposed a Pixel value difference (PVD) and simple least significant bits scheme are used to achieve adaptive least significant bits data embedding. In pixel value differencing (PVD) where the size of the hidden data bits can be estimated by difference between the two consecutive pixels in cover image using simple relationship between two pixels. This method generally provides a good imperceptibility by calculating the difference of two consecutive pixels which determine the depth of the embedded bits. The proposed method hides large and adaptive k-LSB substitution at edge area of image and PVD for smooth region of image. So in this way the technique provide both larger capacity and high visual

quality according to experimental results. However, their algorithm is complex due to adaptive (k) generation for substitution of LSB.

K.-H. Jung et al [10] proposed a method of Multi-Pixel Differencing (MPD) which used more than two pixels to estimate smoothness of each pixel for data embedding and it calculate sum of difference value of four pixels block. For small difference value it uses the LSB otherwise for high difference value it uses MPD method for data embedding. In this method the experimental dataset is too limited.

In [11] authors proposed another pixel value differencing method, it used the three pixels for data embedding near the target pixel. Also it uses simple k-bit LSB method for secret data embedding where number of k-bit is estimated by near three pixels with high difference value. To retain better visual quality and high capacity it simply uses optimal pixel adjustment method on target pixels. In this method the histogram of stego-image and cover-image is almost same, but dataset for experiments are too small.

W. J. Chen et al[12] introduced a high capacity of hidden data utilizing the LSB and hybrid edge detection scheme. For edge computation two types of canny and fuzzy edges detection method applied and simple LSB substitution is used to embed the hidden data. This scheme is successful to embed data with higher peak signal to noise ratio (PSNR) with normal LSB based embedding. This method is tested on limited images dataset.

Madhu et al., in [13] proposed an image steganography method, based on LSB substitution and selection of random pixel of required image area. Also it is target to improve the security where password is added by LSB of pixels. It generates the random numbers and selects the region of interest where secret message has to be hidden. The limitation of this method it is not considers any type of perceptual transparency.

III. PROPOSED ALGORITHM

In this section the methodology of proposed method is given as following:

a. Arnold Transform and LSB Algorithms

Before embedding, the secret message is implemented for block transformation using the Arnold image transformation. The Arnold image transformation is defined as the point (x, y) in the unit square transforms into the other point (x', y') [14]:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \pmod{N} \quad (1)$$

Where, (x', y') {0,1,2,3...N-1} are pixel coordinates of the secret image, (x,y) is the transformed position of (x', y') and N is the order number of image matrix. Suppose the secret image has iterated for K iterations we got the "chaotic" secret image, so K can be saved as a key1. Figure1 show an example of Arnold transform with four iterations.

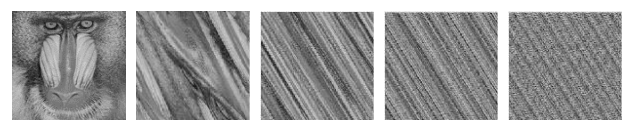


Fig.1. Example of Arnold Transform. The first image is the secret image, which has been encrypted with four iterations separately as shown.

After the encryption step the secret will hide in cover image by using Least Significant Bit (LSB) Replacement. This is the simplest of the steganography methods based in the use of LSB, and therefore the most vulnerable. The embedding process consists of the sequential replacement of each Least Significant Bit (LSB-1-2) of the image pixel for the bit-stream of secret image by bit-or function. The extracting process also consists of sequential extracting for bit-stream of secret image by concatenation method. For its simplicity, this method can camouflage a great volume of information. Figure 2 show the diagram of the proposed method and the following algorithm steps illustrate how the proposed method works.

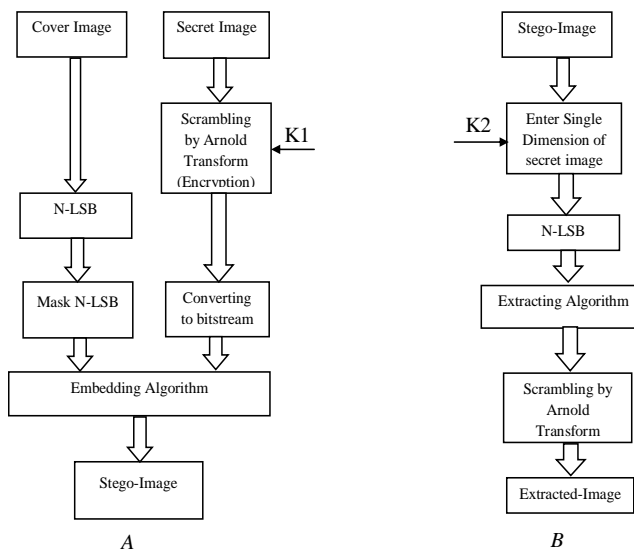


Fig.2. Block Diagram of Proposed Method: (A) Embedding Algorithm, (B) Extracting Algorithm.

In this figure, we have two keys K1 is a key of Arnold transform algorithm which will be the number of iterations. Moreover, K2 is a key of proposed LSB algorithm, when you want to extract the secret image you must enter single dimension of the secret image. By this number of dimension you can extract the right secret image. So, it will be another key and saved as K2.

b. Encryption and Embedding Algorithm

1. Read Cover image and Secret image.
2. Enter the number of iterations for encrypt the secret image (K1).
3. Apply Image scrambling Algorithm using Arnold Transform Method.
4. Enter the N bit plane that will be hide in.
5. Mask N-LSB of image pixel in cover.
6. Convert Secret image to bit stream.

7. If size of bits of secret image bigger than total size of bit space insert-able show error message and read another secret image else continue.
8. Hide first N bit of bit stream in Masked Last N of covered image.
9. Repeat step 6 until all the bits of bit stream embedded.
10. Then create stego image.

c. Decryption and Extracting Algorithm

1. Read Stego image.
2. Enter the single dimension of the secret image (K2).
3. Enter the N-LSB which is hidden in.
4. Find the Length of embedded bitstream.
5. Create a new bit stream.
6. Convert Stego image from decimal to binary.
7. Concatenate the new bit stream and Last N bit of binary Stego image.
8. Convert the new bitstream to array.
9. Then retrieval scrambled image.
10. Enter the number of iterations for decrypt the secret image (K1).
11. Apply Image scrambling Algorithm using Arnold Transform Method.
12. Finally, recuperation the secret image.

IV. RESULT AND DISCUSSION

In Steganography, technique Peak Signal-to-Noise Ratio (PSNR), Normalized Correlation (NC) are standard measures used in order to test the quality of the stego images. PSNR used to evaluate the imperceptibility of the Stego-image, the maximum value is (100) and the minimum value is (0), whenever the bigger the better. It can be found in equation (2). MSE is the Mean Square Error. For imperceptible hiding, the stego-image should look as similar as the cover-image, whenever was the youngest, the better. It can be found in equation (3). In this section, some experiments are performed to demonstrate the efficiency of the proposed method without and with attack. Before the embedding process, the secret image was firstly encrypted using Arnold transform algorithm. Three of 8-bit grayscale images of size 512*512 used as cover and shown in Figure 3.

$$MSE = \frac{1}{M \times N} \sum_{i=0}^M \sum_{j=0}^N (I_B(i, j) - I_H(i, j))^2 \quad (2)$$

$$PSNR = 10 \times \log_{10} \left(\frac{255^2}{MSE} \right) (db) \quad (3)$$



Fig. 3. Cover Images: (A) Baboon (B) Lena (C) Airplane

a. Results without Attack

In this section, the proposed method has been tested in three experiments without attack by taking three standard



512*512 gray scale images (Baboon, Lena and Airplane) as cover and three secret images with different lengths. Figure 4 used "Baboon.bmp" as a Cover image and "Flower.bmp" as Secret image with size (128*128). And Figure 5 used "Lena.bmp" as Cover and "Baboon.bmp" as Secret image with size (192*192). Also Figure 6 used "Airplane.bmp" as cover image and "Lena.bmp" as Secret image with size (256*256).

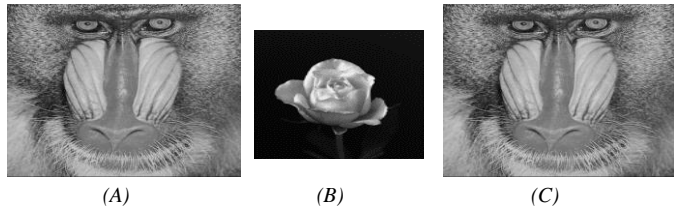


Fig.4. First Embedding Experiment: (A) Cover image, (B) Secret Image, (C) Stego-Image

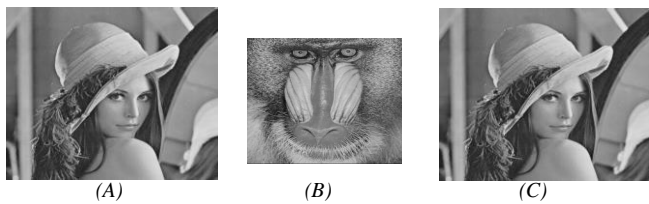


Fig 5. Second Embedding Experiment: (A) Cover image, (B) Secret Image, (C) Stego-Image.



Fig.6. Third Embedding Experiment: (A) Cover image, (B) Secret Image, (C) Stego-Image.

b. -PSNR

The results that are obtained from these experiments are recorded and compared them with another method in [44]. It can be summarized in the following table:

TABLE1. COMPARATIVE PERFORMANCE OF MSE, PSNR WITHOUT ATTACK.

Methods	Cover Image (512*512)	Number of Hidden Bits	PSNR
LSB SM	Lena	164538	38.56
	Baboon	298413	48.18
Proposed Method	Lena	294912	43.60
	Baboon	131072	51.13
	Airplane	524288	44.29

As shown in the table the proposed method has been compared with another method, which labeled (LSB SM) with different capacity for the embedding. Proposed Method has three experiments, while another method has two experiments.

Here PSNR used for indicate the preference. The following histogram will shows the different rate of the two methods.

- Histogram of PSNR

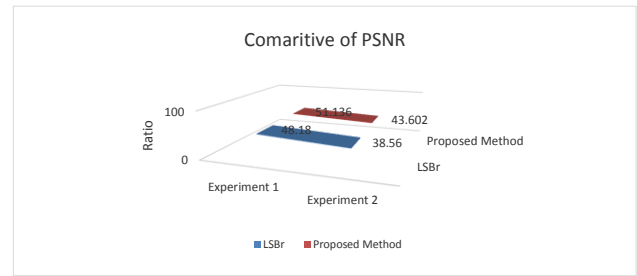


Fig .7. Comparative of PSNR

c. NC

The results of NC that obtained from secret images shown in Figures 8, 9 and10 below and recorded in the next table:

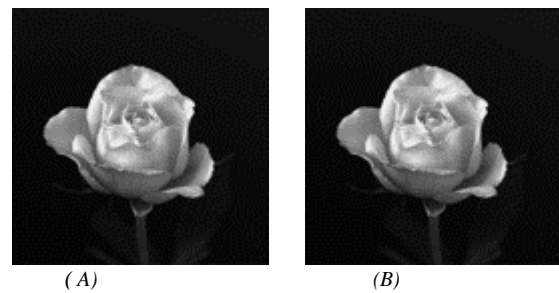


Fig .8. First NC without Attack: (A) Secret Image (128*128), (B) Extracted Image (128*128).

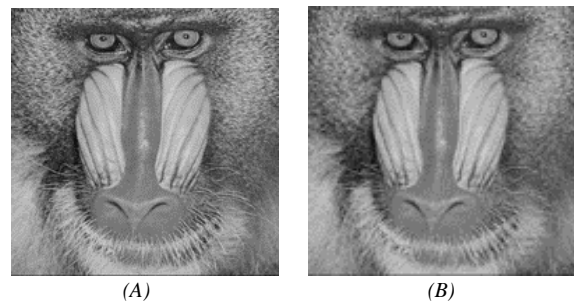


Figure (9): Second NC without Attack: (A) Secret Image (192*192), (B) Extracted Image (192*192).



Fig.10. Third NC without Attack: (A) Secret Image (256*256), (B) Extracted Image (256*256).

TABLE2.RESULT OF NC WITHOUT ATTACK

Secret Image	Extracted Image	NC
Flower (128*128)	Flower (128*128)	1
Baboon (192*192)	Baboon (192*192)	1
Lena (256*256)	Lena (256*256)	1

The table shows the results that are similar NC values which obtained from the different experiments cover-images in Figures 8, 9 and 10, which graphed in the following histogram.

- Histogram of NC

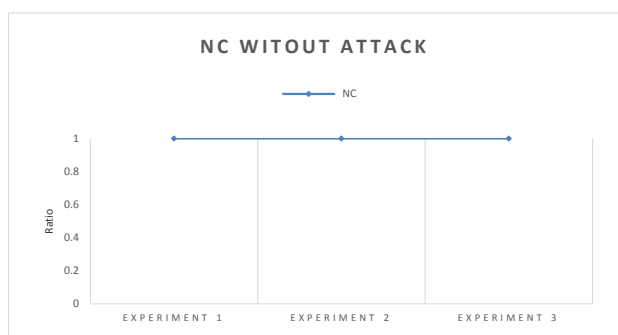


Fig.11. Histogram of NC without Attack.

V. CONCLUSION

The proposed method described in this paper helps to successfully hide the secret image into the cover image, with minimum distortion made to the cover image. First the secret image has been scrambled using Arnold algorithm and embedded to the cover image by using LSB algorithm. This method is essential for construction of accurate targeted and blind Steganalysis methods for BMP images. With using Arnold transform the proposed method will be more secure. The main features of the proposed method are imperceptibility and security. The limitations of the proposed method is slow with extracting algorithm when using large size image Higher than 160*160. Also can not used images which have different dimensions, only can use square array of image which have the same dimensions.

REFERENCES

[1] Cole, E. "Hiding in Plain Sight: Steganography and the Art of Covert Communication", Indiana, John Wiley & Sons Inc, (2003)
 [2] Whitman, M.E. & Mattord, H.J., "Principles of information security". Thomson course technology, 2007.
 [3] Yinglan Fang, Lin Tian "An Improved Blind Watermarking Algorithm for Image Based on DWT Domain", Journal of Theoretical and Applied Information Technology, 15th November 2012. Vol. 45 No.1.

[4] M. Mahdavi*, Sh. Samavi*, N. Zaker** and M. Modarres-Hashemi*, "Steganalysis Method for LSB Replacement Based on Local Gradient of Image Histogram", Iranian Journal of Electrical & Electronic Engineering, Vol. 4, No. 3, July 2008.
 [5] Chang, C.C., Tseng, H.W.: "A Steganographic method for digital images using side match", Pattern Recognition Letters 25, 1431–1437 (2004).
 [6] Wu, H.C, Wu, N.I., Tsai, C.S., Hwang, M.S. "An Image Steganographic Scheme Based on Pixel-Value Differencing and LSB Replacement Methods", IEEE Proceedings of visual image signal Process, November 7, 2004.
 [7] Y. K. Jain and R. R. Ahirwal, "A Novel Image Steganography Method With Adaptive Number of Least Significant Bits Modification Based on Private Stego-Keys", International Journal of Computer Science and Security (IJCSS), vol. 4, (2010) March 1.
 [8] H. Yang, X. Sun and G. Sun, "A High-Capacity Image Data Hiding Scheme Using Adaptive LSB Substitution", Journal: Radioengineering, vol. 18, no. 4, (2009), pp. 509-516.
 [9] C.-H. Yang, C.-Y. Weng, S.-J. Wang, Member, IEEE and H.-M. Sun, "Adaptive Data Hiding in Edge Areas of Images with Spatial LSB Domain Systems", IEEE Transactions on Information Forensics and Security, vol. 3, no. (2008) September 3, pp. 488-497.
 [10] K.-H. Jung, K.-J. Ha and K.-Y. Yoo, "Image data hiding method based on multi-pixel differencing and LSB substitution methods", Proc. 2008 International Conference on Convergence and Hybrid Information Technology (ICHIT '08), Daejeon (Korea), (2008) August 28-30, pp. 355-358.
 [11] H. Zhang, G. Geng and C. Xiong, "Image Steganography Using Pixel-Value Differencing", Electronic Commerce and Security, ISECS '09. Second International Symposium on (2009) May.
 [12] W. J. Chen, C. C. Chang and T. H. N. Le, "High Payload Steganography Mechanism Using Hybrid Edge Detector", Expert Systems with Applications (ESWA 2010), vol. 37, pp. 3292-3301, (2010) April 4.
 [13] V. Madhu Viswanatham and J. Manikonda, "A Novel Technique for Embedding Data in Spatial Domain", International Journal on Computer Science and Engineering, IJCSE, vol. 2, (2010).
 [14] Compressed image file formats: JPEG, PNG, GIF, XBM, BMP / by John Miano, First printing, July (1999), Copyright© 1999 by the ACM Press.

Finger-Knuckle-Print identification System Using Hidden Markov Model and Discret Cosine Transform

Abdallah Meraoumia¹, Salim Chitroub², Ahmed Bouridane³

¹Univ Ouargla, Fac. des nouvelles technologies de l'information et de la communication,
Lab. Génie Electrique, Ouargla 30 000, Algeria

²Laboratory of Intelligent and communication Systems Engineering (LISIC), Electronics and Computer
Science Faculty, USTHB. P.O. box 32, El Alia, Bab Ezzouar, 16111, Algiers, ALGERIA

³Department of Computer Science and Digital Technologies, Northumbria University Newcastle,
Pandon Building, Newcastle upon Tyne NE2 1XE, UK.

¹Ameraoumia@gmail.com, ²S_chitroub@hotmail.com, ³Bouridane@qub.ac.uk

Abstract—Automatic personal identification from their physical and behavioral traits, called biometrics technologies, is now needed in many fields such as: surveillance systems, security systems, physical buildings and many more applications. In this paper, we propose an efficient online personal identification system based on Finger-Knuckle-Print (FKP) using the Hidden Markov Model (HMM) and two-dimensional Block based Discrete Cosine Transform (2D-BDCT). In this study, a segmented FKP is firstly divided into non-overlapping and equal-sized blocks, and then, applies the 2D-BDCT over each block. By using zigzag scan order (starting at the top-left) each transform block is reordered to produce the feature vector. Subsequently, we use the HMM for modeling the feature vector of each FKP. Finally, Log-likelihood scores are used for FKP matching. Our experimental results show the effectiveness and reliability of the proposed approach, which brings both high identification and accuracy rate.

Keywords— *Biometrics; identification; Finger-Knuckle-Print; 2D-BDCT, HMM.*

I. INTRODUCTION

Traditionally, identification strategies are based on something we know, e.g., a password or a personal identification number (PIN), or something we own, e.g., a card, or a key. Unfortunately, passwords can be guessed by an intruder; cards can be stolen or lost. Biometrics, which deals with identification of individuals based on their physical or behavioral features, has been emerging as an effective identification technology to achieve accurate and reliable identification results. The biometrics has significant advantages over traditional identification techniques due to biometric characteristics of an individual are not transferable and unique for every person and are not stolen or broken [1].

Currently, a number of biometrics-based technologies have been developed and hand-based person identification is one of these technologies. This technology provides a reliable, low cost and user-friendly viable solution for a range of access control applications. In contrast to other modalities, like face and iris, hand biometry offers some advantages [2]. First, data acquisition is economical via commercial low-resolution cameras, and its processing is relatively simple. Second, hand based access systems are very suitable for several usages. Finally, hand features are more stable over time and are not susceptible to major changes. Some features related to a human hand are relatively invariant and distinctive to an individual. Among these features, Finger-Knuckle-Print (FKP) is one biometric that has been systematically used to make identification for last years. FKP identification is a biometric

technology which recognizes a person based on his/her finger knuckle pattern. The rich texture information of FKP offers one of the powerful means in personal identification [3].

An important issue in FKP identification is to extract FKP features that can discriminate an individual from the other. Based on texture analysis, our biometric identification system used the 2D-BDCT for features extracted from FKP images. In this method, a FKP is firstly divided into non-overlapping and equalized blocks, and then, applies the 2D discrete cosine transform over each block. By using zigzag scan order each transform block is reordered to produce the feature vector, then concatenated all vectors for produce an observation vector. Subsequently, we use the HMM for modeling the observation vector of each FKP. Finally, Log-likelihood scores are used for FKP matching. In this work, a series of experiments were carried out using a FKP database. To evaluate the efficiency of this technique, the experiments were designed as follow: the performances under different finger types were compared to each other, in order to determine the best finger type at which the FKP identification system performs. However, because our database contains FKPs from four types of fingers, an ideal FKP identification system should be based on the fusion of these fingers at different fusion levels.

The rest of the paper is organized as follows. The proposed scheme of the unimodal biometric system is presented in section 2. Feature extraction and modeling process are discussed in section 3. This section including also an overview of 2DBDCT and the HMM-modeling. The experimental results, prior to

fusion and after fusion, are given and commented in section 5. Finally, the conclusions and further works are presented in sections 6.

II. SYSTEM OVERVIEW

The proposed system consists of preprocessing, feature extraction, matching and decision stages. To enroll into the system database and modeling, the user has to provide a set of training FKP images. Typically, an observation vector is extracted from each finger which describes certain characteristics of the FKP images using Discrete Cosine Transform technique and modeling using Hidden Markov Model. Finally, the models parameters are stored as references models. For identification, the same observation vectors are extracted from the test FKP images and the log-likelihood is computed using all of models references in the database. Our database contains FKPs from four types of fingers, for this reason, each FKP modalities are used as inputs of the matcher modules (subsystem). For the multimodal system, each subsystem compute its own matching score and these individual scores are finally combined into a total score (using fusion at the matching score level), which is used by the decision module. We have also tried the various image fusion rulers and various feature extraction fusion rulers to choose the best one for FKPs classification.

III. FEATURE EXTRACTION AND MODELING

A. 2D Block based discrete cosine transform

Discrete Cosine Transform (DCT) is a powerful transform to extract proper features for FKP identification. The DCT is the most widely used transform in image processing algorithms, such as image/video compression and pattern recognition. Its popularity is due mainly to the fact that it achieves a good data compaction, that is, it concentrates the information content in a relatively few transform coefficients [5]. In the 2D-BDCT formulation, the input image is first divided into, $\eta_1 \times \eta_2$ blocks, and the 2D-DCT of each block is determined. The 2D-DCT can be obtained by performing a 1D-DCT on the columns and a 1D-DCT on the rows. Given an image, where, represent their size, the DCT coefficients of the spatial block are then determined by the following formula:

$$F_{ij} = C(v)C(u) \sum_{m=0}^{M-1} \sum_{n=0}^{M-1} f_{ij}(n, m) \psi(n, m, u, v) \quad (1)$$

$$\psi(n, m, u, v) = \cos \left[\frac{(2n+1)u\pi}{2M} \right] \cos \left[\frac{(2m+1)v\pi}{2M} \right] \quad (2)$$

$u, v = 0, 1, \dots, M-1, i = 1, \dots, \eta_1, j = 1, \dots, \eta_2$ with $\eta_1 = H/M, \eta_2 = W/M$ and $F_{ij}(u, v)$ are the DCT coefficients of the B_{ij} block, $f_{ij}(n, m)$ is the luminance value of the pixel (n, m) of the B_{ij} block, $H \times W$ are the dimensions of the image, and

$$C(u) = \begin{cases} \frac{1}{\sqrt{2}} & \text{if } u = 0 \\ 1 & \text{if } u \neq 0 \end{cases} \quad (3)$$

After transformation process, if $M = 8$, there will be 64 DCT coefficients contained within each transformed block, where the coefficient at the top-left is called DC ($F_{ij}(0, 0)$) coefficient and the rest is called AC coefficients.

B. Observation vector

The block-based approach partitions the input image, with size $H \times W$, when $H = 220$ and $W = 110$, into small non-overlapped blocks; each of them is then mapped into a block of coefficients via the 2D-DCT. Most popular block size is commonly set to $M \times N$ with $M=8$. The number of blocks extracted from each FKP image equals to:

$$\eta = \lceil \eta_1 \rceil * \lceil \eta_2 \rceil = \left\lfloor \frac{220}{8} \right\rfloor * \left\lfloor \frac{110}{8} \right\rfloor = 351 \text{ blocks} \quad (4)$$

Then, we form a feature vector from the 2D-DCT coefficients of each image block. The 2D-DCT concentrates the information content in a relatively few transform coefficients top-left zone of block, for this, the coefficients, where the information is concentrated, tend to be grouped together at the start of the reordered array. Thus, a suitable scan order is a zigzag starting from the DC (top-left) coefficient. Starting with the DC coefficient, each coefficient is copied into a one-dimensional array. So, each block can be represented by a vector of coefficients:

$$O_{ij} = [F_{ij}(0,0) \ F_{ij}(0,1) \ F_{ij}(1,0) \ \dots \ F_{ij}(U, V)]^T \quad (5)$$

U, V are chosen as well as the identification rate was maximum. Thus, $U, V \in [0 .. 7]$ and the size of O_{ij} is τ with $\tau \in [1 .. 64]$. Finally, the results o_{ij} of a blocks image are combined in the single template as follows:

$$V_{obs} = [O_{11} \ O_{12} \ O_{13} \ \dots \ O_{\eta_1 \eta_2}] \quad (6)$$

where the size of resulting observation vector is $[\tau, \eta]$.

C. Hidden Markov model

A hidden Markov model is a collection of finite states connected by transitions. Each state is characterized by two sets of probabilities [6]: a transition probability and either a discrete output probability distribution or continuous output probability density function which, given the state, defines the condition probability of emitting each output symbol from a finite alphabet or a continuous random vector. An HMM can be written in a compact notation $\lambda = (A, B, \pi)$ to represent the complete parameter set of the model, where A , B , and π represents, respectively, state transition probability distribution, probability distribution of observation symbols and initial state distribution. Finally, forward backward recursive algorithm, Baum-Welch algorithm and Viterbi algorithm are used to solve evaluating, training, and decoding, respectively [7].

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. Experimental database

We experimented our method on Hong Kong polytechnic university (PolyU) FKP Database [8]. The database has a total of 7920 images obtained from 165 persons. This database

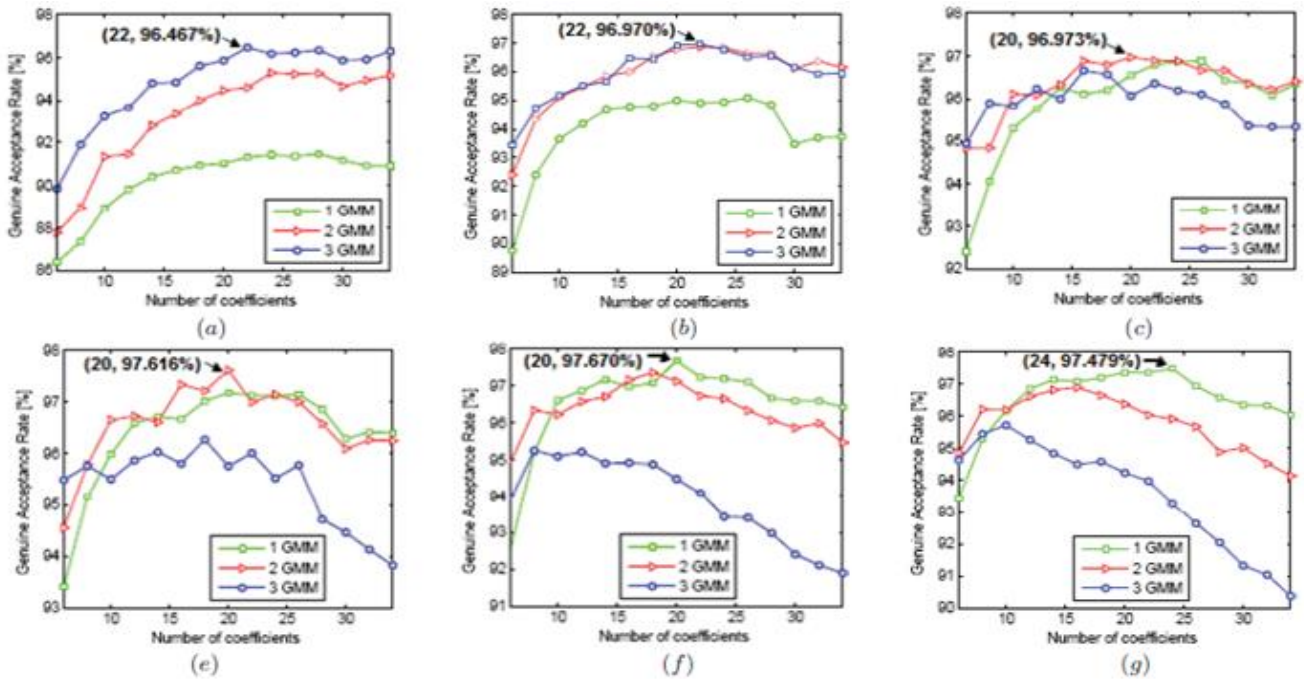


Fig. 1. System performance under different stats number. (a) One state, (b) Two states, (c) Three states, (d) Four states, (e) Five states and (f) Six states.

including 125 males and 40 females. Among them, 143 subjects are 20~30 years old and the others are 30~50 years old. These images are collected in two separate sessions. In each session, the subject was asked to provide 6 images for each of Left Index Fingers (LIF), Left Middle Fingers (LMF), Right Index Fingers (RIF) and Right Middle Fingers (RMF). Therefore, 48 images were collected from each subject.

B. Selecting 2D-BDCT coefficients and HMM parameters

A series of experiments were carried out using the FKP database to selection the best number of 2D-BDCT coefficients and the HMM parameters (number of states and number of Gaussian Mixture Model (GMM)), This is carried out by comparing all states and k -GMM, with $k=1$ to 6 and $s=1$ to 3, for several 2D-BDCT coefficients and finding the number of states and GMM that gives the best identification rate. The problem we address is as follows: we want chosen the number of 2D-BDCT coefficients, the states and their k -GMM such that the Genuine Acceptance Rate (GAR) is maximized. Thus, the 2D-BDCT coefficients reflect the compact energy of different frequencies. Most of the higher frequency coefficients are small and they become negligible, as result, the features derived from the 2D-BDCT computation is limited to an array of summed spectral energies within a block in frequency domain [9]. In Fig. 1, we plot the system performance as a function of the number of 2D-BDCT coefficients selection in each block for various numbers of GMM and various numbers of states in the HMM. The reason Fig. 1 was generated was to show how the number of 2D-BDCT coefficients, numbers of states in the HMM and GMM used might have an effect on the performance of our system. We observe that the identification accuracy becomes very high at certain coefficients, where it actually exceeds 96 % and slight decrease in identification accuracy as we go to higher

numbers of coefficients. Also, note that only 20 coefficients with $M = 5$ states and $k = 1$ GMM are enough to achieve good accuracy (see Fig. 1.(f)).

C. Unimodal identification test results

The goal of this experiment was to evaluate the system performance when we using information from each modality (each finger). For this, we found the performance under different modalities (LIF, LMF, RIF, and RMF). By adjusting the matching threshold, a ROC curve, which is a plot of FRR against FAR for all possible thresholds, can be created. For this, the numbers of training and test samples are 495 and 1485. We matched all the 1485 FKP images (test) with each other to obtain 245025 distances. Thus, we have a total of 1485 genuine matching and the remaining, 243540, impostor matching. Fig. 2.(a) compares the performance of the system for varying fingers. The experimental results indicate that the LIFs perform better than the LMFs, RIFs and RMFs in terms of Equal Error Rate (EER) (2.282 %). Therefore, the system can achieve higher accuracy at the LIFs compared with the other fingers of a person. The results expressed as a False Acceptance Rate (FAR) and False Rejection Rate (FRR) depending on the threshold and the distance distributions of genuine and impostor matching's obtained by the proposed scheme, if the LIF is used, are plotted in Fig. 2.(b) and Fig. 2.(c), respectively. Finally, the system was tested with different thresholds and the results are shown in Table. 1.

D. Multimodal identification test results

The goal of this experiment was to investigate the systems performance when we fuse information from some fingers of a person. In fact, in such a case the system works as a kind of

TABLE I. OPEN SET IDENTIFICATION TEST RESULTS IN THE CASE OF SINGLE BIOMETRIC

DATABASE	LEFT INDEX FINGER			LEFT MIDDLE FINGER			RIGHT INDEX FINGER			RIGHT MIDDLE FINGER		
	T_o	FAR	FRR	T_o	FAR	FRR	T_o	FAR	FRR	T_o	FAR	FRR
165 Persons	0.9500	7.125	0.889	0.9500	6.712	1.481	0.9300	6.549	1.704	0.9100	7.719	0.963
	0.9699	2.282	2.282	0.9669	2.754	2.754	0.9502	2.998	2.998	0.9449	2.297	2.296
	0.9900	0.352	9.333	0.9900	0.355	9.185	0.9800	0.573	8.444	0.9800	0.350	6.593

TABLE II. OPEN SET IDENTIFICATION TEST RESULTS IN THE CASE OF FUSION AT IMAGE LEVEL

COMBINATION	DWT		PCA		LAPLACIAN		GRADIANT		CONTRAST	
	T_o	EER	T_o	EER	T_o	EER	T_o	EER	T_o	EER
LIF-LMF	0.9797	3.015	0.9774	3.646	0.9775	2.856	0.9802	2.558	0.9780	2.587
LIF-RIF	0.9738	3.165	0.9671	4.770	0.9727	2.786	0.9752	2.505	0.9728	2.517
LMF-RMF	0.9709	2.788	0.9655	4.088	0.9688	2.421	0.9708	2.440	0.9690	2.146
RIF-RMF	0.9588	2.905	0.9540	3.034	0.9586	2.522	0.9605	2.294	0.9582	2.458
All Fingers	0.9800	3.086	0.9658	3.254	0.9760	3.182	0.9790	2.646	0.9760	2.992

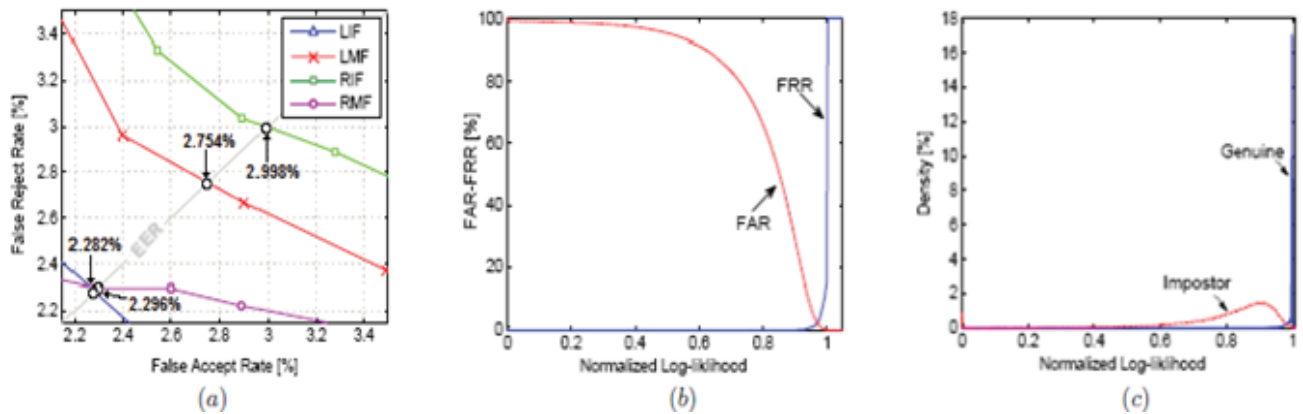


Fig. 2. Unimodal system performance. (a) The ROC curves for all finger types, (b) The dependency of the FAR and the FRR on the value of the threshold (LIF modality) and (c) The genuine and impostor distribution (LIF modality).

multimodal system with a single biometric trait but multiple units. Therefore, information presented by different biometrics (finger types) is fused to make the system efficient.

1) *Fusion at image level:* Image fusion is the process by which two or more images are combined into a single image. For that, a series of experiments were carried out using the FKP database to selection the best combination and fusion technique (DWT, PCA, LAPLACIAN, GRADIANT and CONTRAST) [10, 11] that maximize the GAR. However, in order to see the performance of the identification system, we usually give, in Table 2, the results for all the fusion techniques and the possible combinations. Thus, the result suggests that the fusion of LMF and RMF with CONTRAST technique has performed better than other (EER = 2.146 % and= 0.9690).

1) *Fusion at feature level:* We also investigated the integration of multiple biometric modalities at the representation level. The data obtained from each biometric modality (LIF, LMF, RIF and RMF) is used to compute a feature vector. The idea of fusion at the feature extraction level is to concatenate the feature vectors of different biometrics (different fingers). The new observation vector has a higher dimensionality and represents a person's identity in a different

feature space. Several fusion techniques has been proposed by various researchers. To find the better of the all fusion techniques, with the lowest EER, table showing the results were generated (see Table 3). This Table shows that the LMF and RMF combination with HORIZONTAL technique offers better results (EER = 1.126 % and = 0.9618).

2) *Fusion at matching score level:* In our system the individual matching scores are combined to generate a single scalar score, which is then used to make the final decision. During the system design we experimented five different fusion schemes [12]: Sum-score (SUM), Sum-weighting-score (WHT), Min-score (MIN), Max-score (MAX) and Mul-score (MUL). Table 4 provides the performance of the identification system. From Table 4, it is clear that our identification system achieves a best performance when using all finger with Sum rule fusion (EER = 0.269 % and = 0.9676).

In Fig. 4.(a), we compare the performance of unimodal and multimodal system. The results show the benefits of using the multimodal system with matching score level fusion. Finally, the results expressed as a FAR and FRR depending on the threshold and the distance distributions of genuine and imposter matching's obtained by the proposed scheme, if the all fingers

are fused in the case of matching score level by SUM rule, are plotted in Fig. 4.(b) and Fig. 4.(c), respectively.

V. CONCLUSION AND FURTHER WORK

This paper proposes an efficiency scheme for FKP identification using the HMM and 2D-BDCT. Firstly the ROI is divided into non-overlapping and equal-sized blocks, and then, applies the DCT over each block to produce the feature vector. Subsequently, we use the HMM for modeling the feature vector of each FKP. Finally, Log-likelihood scores are used for the matching process. The proposed scheme is validated for their efficacy on PolyU FKP database of 165

[1] Arun A. Ross, K. Nandakumar and A. K. Jain, “Handbook of Multibiometrics”, in Springer Science+Business Media, LLC, New York, 2006.
 [2] Ajay Kumar and David Zhang, “Improving Biometric Authentication Performance From the User Quality”, in *IEEE transactions on instrumentation and measurement*, vol. 59, no. 3, march 2010.
 [3] R. Zhao, K. Li, M. Liu, X. Sun, “A Novel Approach of Personal Identification Based on Single Knuckleprint Image”, in *AsiaPacific Conference on Information Processing, APCIP*, 2009.
 [4] L. Zhang, L. Zhang, D. Zhang, “Finger-knuckle-print: a new biometric identifier”, in: *Proceedings of the ICIP09*, 2009.

TABLE III. OPEN SET IDENTIFICATION TEST RESULTS IN THE CASE OF FUSION AT FEATURE LEVEL

COMBINATION	HORIZONTAL		ROW		COLUMN		VERTICAL		MEAN	
	T_o	EER	T_o	EER	T_o	EER	T_o	EER	T_o	EER
LIF-LMF	0.9720	4.682	0.9806	1.367	0.9786	1.316	0.9607	1.646	0.9804	2.472
LIF-RIF	0.9798	1.230	0.9562	1.305	0.9741	1.564	0.9561	1.305	0.9722	3.142
LMF-RMF	0.9618	1.126	0.9619	1.226	0.9717	1.679	0.9619	1.224	0.9771	3.090
RIF-RMF	0.9495	4.748	0.9959	2.772	0.9471	2.991	0.9959	2.772	0.9702	3.117
All Fingers	0.9698	7.175	0.9880	1.308	0.9727	1.695	0.9880	1.307	0.9771	2.953

TABLE IV. OPEN SET IDENTIFICATION TEST RESULTS IN THE CASE OF FUSION AT MATCHING SCORE LEVEL

COMBINATION	SUM		WHT		MIN		MAX		MUL	
	T_o	EER	T_o	EER	T_o	EER	T_o	EER	T_o	EER
LIF-LMF	0.9736	0.822	0.9736	0.826	0.9857	1.097	0.9631	1.269	0.9489	0.806
LIF-RIF	0.9649	0.984	0.9667	0.889	0.9807	1.369	0.9530	1.407	0.9317	1.004
LMF-RMF	0.9542	1.116	0.9667	0.889	0.9795	1.183	0.9424	1.492	0.9163	1.100
RIF-RMF	0.9543	0.815	0.9519	0.889	0.9758	0.983	0.9404	1.259	0.9143	0.872
All Fingers	0.9676	0.269	0.9682	0.278	0.9990	0.450	0.9330	1.255	0.8829	0.296

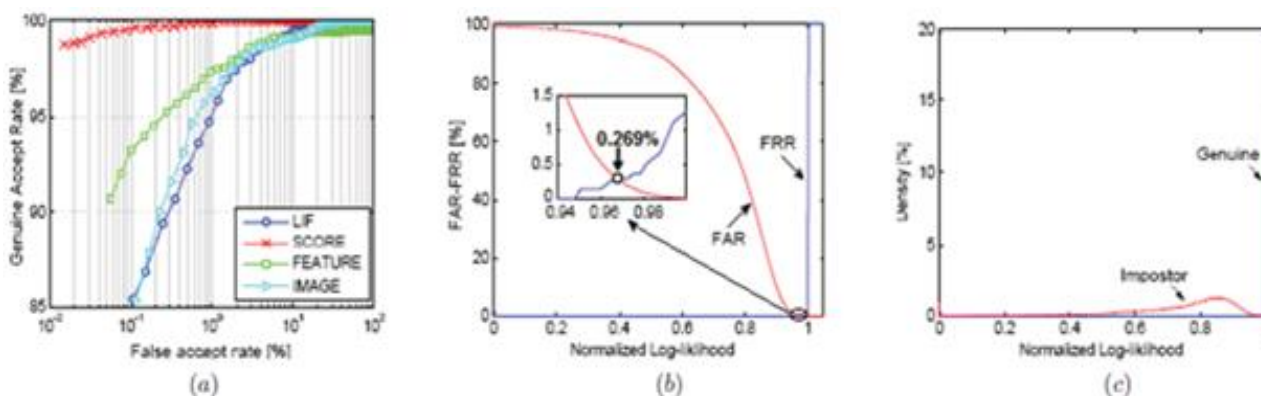


Fig. 3. Multimodal system performance in the case of fusion at matching score level (all fingers) with SUM rule. (a) The comparison between the unimodal and multimodal systems, (b) The dependency of the FAR and the FRR on the value of the threshold and (c) The genuine and impostor distribution.

users. Our experimental results experimental results indicate that the proposed system has a good capability to identify a person’s identity. Our future work will focus on the performance evaluation using other fusion level (*decision*), and combining both FKP and palmprint to get security system with high accuracy.

REFERENCES

[5] Ahmed N, Natarajan T, and Rao K, “Discrete cosine transform”, in *IEEE Trans, on Computers*, 23(1):9093.
 [6] H. Uguz, A. Arslan, “A biomedical system based on hidden Markov model for diagnosis of the heart valve diseases”, in *Pattern Recognition Letters* 28, 2007, pp: 395-404.
 [7] Fabien Cardinaux, Conrad Sanderson, Samy Bengio, “Face Verification Using Adapted Generative Models”, in *6th IEEE Int. Conf. Automatic Face and Gesture Recognition*, Seoul, 2004, pp. 825-830.

- [8] PolyU Finger-Knuckle-Print Database.
<http://www4.comp.polyu.edu.hk/~biometrics/ FKP.htm>.
- [9] Ziad M. Hafed and Martin D. Levine, "Face Recognition Using the Discrete Cosine Transform", in *International Journal of Computer Vision* 43(3) , 2001, pp: 167-188.
- [10] Akansu, A. N. and Haddad, R. A, "Multiresolution signal decomposition", *Academic Press*, New York, 1992.
- [11] Mallat, S. G, "A theory for multiresolution signal decomposition", in *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 11, No. 7, 1989, pp. 674-693.
- [12] A. Meraoumia, S. Chitroub and A.Bouridane, "Palmprint and Finger-Knuckle-Print for efficient person recognition based on LogGabor filter response", in *International Journal Analog Integrated Circuits and Signal Processing*, Vol 69 (1), pp: 17-27, 2011.

A Suggested Algorithm of Recommender System to Recommend crawled-Web Open Educational Resources to Course Management System

Jamil A. Itmazi

Director of e-Learning center
Palestine Ahliya University (PAU)
Bethlehem, Palestine
j.itmazi@gmail.com

Haytham W. Hijazi

Director of the computer center
Palestine Ahliya University (PAU)
Bethlehem, Palestine
haitham@paluniv.edu.ps

Abstract — The majority of educational institutes and training centers are using some kinds of e-Learning via online platform i.e. Course Management System (CMS), Learning Contents Management System (LCMS). These platforms are somehow fixed to the e-contents of the tutor and teacher, while there are huge Open Educational Resources (OERs) available in the Web and ready for using and sharing. This paper proposes a Recommender System (RS) to recommend automatically OERs to a CMS after crawling them from Web to solve the students "Information Overload" problem arising from searching Web resources. This paper provides background of CMS - LCMS and RS as well as some examples. In addition, it discusses the suitability of main RS approaches to recommend digital resources from Web to support students' needs. Finally, it presents a new proposal of RS algorithm which could automatically recommend suitable digital learning resources to a student in his active course.

Keywords— *Recommender System, Open Educational Resources, Course Management System, Learning Management System*

I. INTRODUCTION

Using web to deliver educational contents is the latest trend in training and education development industry [26]. However, the majority of the current online learning content management systems (LCMSs) are somehow fixed to the e-contents of the tutor and teacher. Moreover, the majority contents of the current LCMSs cannot be customized upon the learner's interests.

A. Open Educational Resources (OERs)

We agree with whom calling Open Educational Resources (OERs) as a learning revolution since we see many big universities adopt OERs. Moreover, many other institutes started OERs Initiatives in the time we see some of OERs and Massive Open Online Courses (MOOCs) gain millions of students. Nowadays there are huge data related to OER ready for using, sharing, customizing. The main question here is how can we benefit from these huge OERs?

Many educational institutes and training centers are using an eLearning platform (LCMSs) to automate the administration of their training events and educational contents. In the same time, the students suffer from the "Information Overload" problem, when they find in the Web thousands of results that are not suitable and not related to their LCMSs courses (eCourse).

Consequently, to solve this problem, the institute needs RS (Recommender System) to present interesting educational resources which relates to students eCourse as well as fit their preferences from the internet via a topical web crawler (i.e., focused spider) that visits the web sites of OERs, crawl its contents, and index them based on the keywords. The retrieved resources could be listed automatically in a particular eCourse upon suitable priorities and ranking upon their importance to the tutor and learner. Figure No. 1 represents our RS system. Open Educational Resources (OERs).

Our web crawler will worth nothing if there are no open resources to crawl. Fortunately there are many educational websites allow the accessing and retrieving of their educational resource especially the OERs websites.

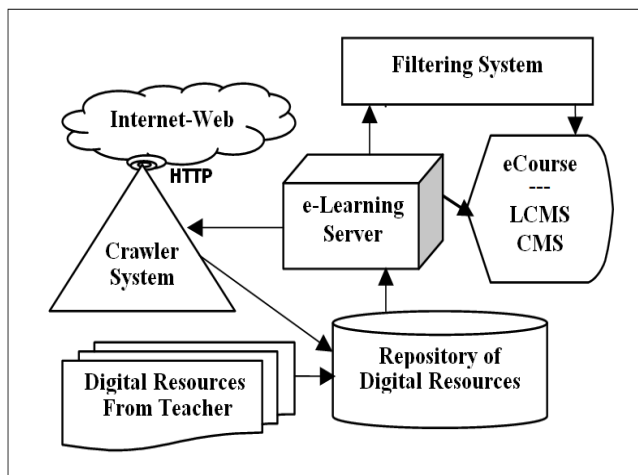


Fig. 1. A general structure of RS system

OERs began in 2001 when Massachusetts Institute of Technology (MIT) first announced to making all of its course materials freely available [16]. Its initiative called MIT OpenCourseWare (OCW) under the Website (<http://ocw.mit.edu>) to publish all of its undergraduate and graduate-level courses materials online, partly free and openly available to anyone, anywhere.

As of March 2014, over 2200 courses were available online [14].

The initiative has inspired a number of other institutions to make their course materials available as OER. Nowadays there are many institutions offer their course materials available online. We can find 40 of them in [6]. More and more could be found in [1]. As an example, the following projects and resources have been selected to illustrate the richness and diversity of the current initiatives in open educational and related resources and practices:

- M.I.T. Open Courseware (OCW), <http://ocw.mit.edu>
- OpenLearn – Open University UK, <http://openlearn.open.ac.uk>
- World Bank - Open Knowledge Repository, <https://openknowledge.worldbank.org>
- Center for Open and Sustainable Learning (COSL) / OpenEd conferences, <http://cosl.usu.edu>
- Commonwealth of Learning – COL's Directory of Open Education Resources, www.col.org
- Connexions (online platform for managing and sharing of open course modules), <http://cnx.org>

- Creative Commons, <http://creativecommons.org>
- Directory of Open Access Journals, www.doaj.org
- Arabic Open programming school, barmaje.com

The term OER has been used to refer to learning materials such as:

- Learning objects (quizzes, animations, interactive maps, timelines, etc.)
- Audio lectures
- Audio-video lectures
- Images
- Sounds and music
- Entire course content
- Collections of journal articles and institutional repositories
- Textbooks

B. Course Management System (CMS)

It called also Learning Content Management System (LCMS) is an eLearning platform which is considered as an important part of eLearning solutions [9]. Moreover, there are some concepts similar to LMS (with a small difference), e.g. LMS (Learning Management System) and Portal Learning.

Generally, CMS is software that automates the administration of training events; it manages the log-in of registered users, manage course catalogs, track learner activities and results, as well as provide reports to management.

The market of CMS, LCMS and LMS is increasing fast, and there are hundreds of them in the market, for example [13] list 599 of them; some of CMS, LCMS and LMS are commercial Software, while others are free Open-Source LMSs. The following list shows some LMSs:

- Commercial LMS: e.g. WebCT <www.WebCT.com> and eCollege <www.ecollege.com>.
- Open-Source LMS: e.g. MOODLE <<http://moodle.org>> and ILIAS <www.ilias.de>.

C. Focused crawler

The e-course content has increasingly become a focus for the learning process. The educational contents' update, however, has become a growing challenge to produce high quality and fresh educational material. Therefore, in this model we implement an effective web crawler to crawl the websites of OERs, e.g. <http://ocw.mit.edu>, download the

recent course-related educational resources, and integrate these updates with the existing course materials.

The web crawler [27, 4, 22, 28] is a web application that crawl websites by taking a list of seed URLs as an input, determine the IP address of the host name, download the corresponding resources and extract the links to continue the process. For example, the courseware Watchdog [23, 20] which is part of the Personalized Access to Distrusted Learning Repositories (PADLR) framework has a focused or a topical web crawler to retrieve learning material from the WWW to be a part the learning materials.

In order to restrict the crawling to a material-relevant process, the crawling is done topically or focused.

II. RECOMMENDATIONS SYSTEMS

Have been widely implemented and accepted in many Internet sectors [5]. We are familiar with recommendations of products (e.g. books, music and movies) and of services (e.g. restaurants, hotels, Web sites); likewise recommendation is not arising from the digital era, but an existing social behavior in daily life. In everyday life, we rely on recommendations from others. [10].

Generally, the internet reaches the Billion Terabytes of data and the Web is still growing faster; as a result, the users suffer from the "Information Overload" problem, when searching the Internet [21]. Fortunately, the aim of RSs in Web applications is to present interesting information that fits the users' tastes and preferences with little effort.

In contrast, sometimes RSs are used to hide special information, and specifically, the aim of RSs in eLearning applications is listing "the closest available learning objects to what the instructor describes as the module's content" [3].

A. Current Usage of RS

RSs have been widely used in many Internet activities. It is worth mentioning some examples of the current actual uses of RS:

- eCommerce: RSs are used "to suggest products to their customers and provide consumers with information to help them decide which products to purchase" [19]. eCommerce leaders like Amazon.com and Netflix have made recommender systems a salient part of their websites [11].
- Web pages: RS is used to solve the "overload problem" in the Internet, when using search engines (e.g. Google, Bing, yahoo) which produce thousands of pages to one researched item; most of them have worthless relation to the researched item or of no interest to the user. Example of search engines which used RS: Mi Yahoo! <http://my.yahoo.com> and Alexa.com.
- Censorship systems: RSs used to protect children from accessing undesirable material on the internet. e.g. cyberpatrol.com, as well as Prevent citizens from exploring some Web sites; which some governments already did.
- Other sectors: Examples:
 - 1) News: e.g. www.lemonde.fr,
 - 2) Encyclopedia: e.g. <http://wikipedia.org>,
 - 3) Software: e.g. <http://download.cnet.com>,
 - 4) Stores: e.g. www.drugstore.com,
 - 5) Tourist information: e.g. www.viamichelin.com,
 - 6) Digital library: e.g. <http://ieeexplore.ieee.org> and <http://citeseerx.ist.psu.edu>.

B. RS and eLearning

eLearning is able to apply RS, which may be used to recommend the most appropriate content to students. In this paper, the focus will be at the use of RS in CMS. Some researchers mentioned the abilities of using RS in eLearning systems in general and CMS in particular. [3, 2, 12, 8, 18].

In the following we introduce the suitability of using RS approaches in the CMS.

III. THE SUITABILITY OF RS APPROACHES

RSs consist of approaches; every approach has its techniques. However, there are many systems that use Hybrid Recommender System, which combines two or more recommendation techniques to gain better performance.

Here, we are going to preview the suitability of the main RS approaches to recommend digital resources from Web to a CMS:

A. Content-Based System

In this type, the resources are selected by having correlation between the content of the resources and the user's preferences. In the context of this research this system could be used within RSs as a primary approach to find the digital resources from Websites, by detecting similarities between the current eCourse attributes (name, keywords, abstract ...etc.) and the OERs attributes.

B. Collaborative Filtering Systems

It recommends items or resources to a target user, based on similar users' preferences, and on the opinions of other users with similar tastes. It employs statistical techniques to find a set of users known as neighbors to the target user, examples: Amazon.com and ebay.com.

This system has some methods to calculate the likeliness from the rating matrix, the suitable one to our RS is Memory-Based Algorithm (also known as k-Nearest Neighbor Method), and because it is suitable to environments where the user preferences have to be updated rapidly.

C. Demographic-Based Filtering

It uses “prior knowledge on demographic information about the users and their opinions for the recommended items as basis for recommendations” [15]. It aims to categorize the user based on personal explicit attributes and make recommendations based on demographic group that a user belongs to, such as (income, age, learning level, or geographical region), or a combination of these clusters/groups. E.g. the Free e-mail suppliers put advertisements based on the user demographic information, like Yahoo as well as Google search engine.

The Demographic-Based Filtering could be used in the process of recommending digital resources as a complementary approach.

D. Rule-Based Filtering

It is filtering information according to set of rules expressing the information filtering policy [24]. These rules may be part of the user or the system profile contents and it may refer to various attributes of the data items. In general, this system is used widely with:

- Censorship: It is useful in the protection domain e.g. the protection of kids from accessing some materials, e.g. Cyberpatrol.com and Cybersitter.com [7].
- Spam Filtering: It is useful to be used against the Spam e-mails, e.g. Spam Assassin <http://spamassassin.apache.org> and MailEssentials <www.gfi.com>.

This system could be used within RSs to filter the recommendations list of digital resources upon some rules of the system and the student.

E. Hybrid Recommender System

“It combines two or more recommendation techniques to gain better performance with fewer of the drawbacks of any individual one”. [17]. Examples of systems [21, 25].

IV. A GENERAL RS PROPOSAL

The suitable RS approach to recommend OERs from Web to CMS will not be a pure one, but it will be a Hybrid, which mixed some of the previous approaches.

The following general RS structure could be suggested (see Figure 2).

We list some consideration of this proposal structure:

- Content-Based System is used as a primary approach because it can give comprehensive, related and sufficient recommendations by using the resources attributes in the recommendation process.
- Collaborative Filtering Systems is not used as a primary approach because this approach becomes useful only after a "critical mass" of opinions, which means less numbers of recommendations or null recommendations.

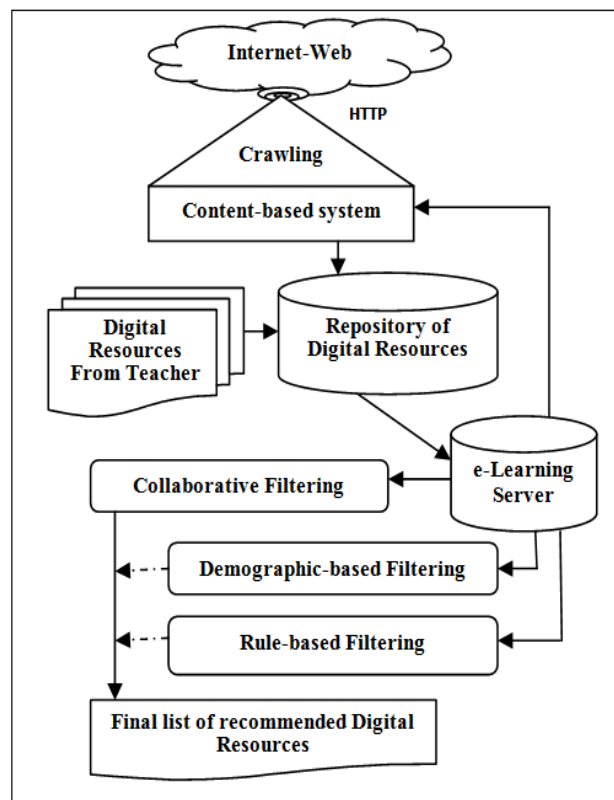


Fig. 2. A general proposal structure of RS algorithm

- Demographic-Based Filtering and Rule-Based Filtering used as complementary approaches, because the demographic information and the rules are not useful to be a primary approach.
- The recommendations will appear at the eCourse window when the student enters his eCourse.

A. The Stage of Crawling, Fetching and Retrieving

In this stage, the Crawler System via Content-Based System will Fetch the admin suggested websites for digital resources and select these resources by detecting similarities between the items of the eCourses in CMS and the items of digital resources in websites, then the Crawler System retrieves them into the Repository of Digital Resources.

The eCourse items include: name, keywords, abstract ...etc. Empirically, as shown in Figure 3, the tutor configures the crawler for his own course -- CSS is the example here. First, the tutor is supposed to enter the seed URL/URLs for his course, and decided the suitable keywords for the course content. Second, he determines the successive crawling frequency. And finally, the type of fetched content is selected.

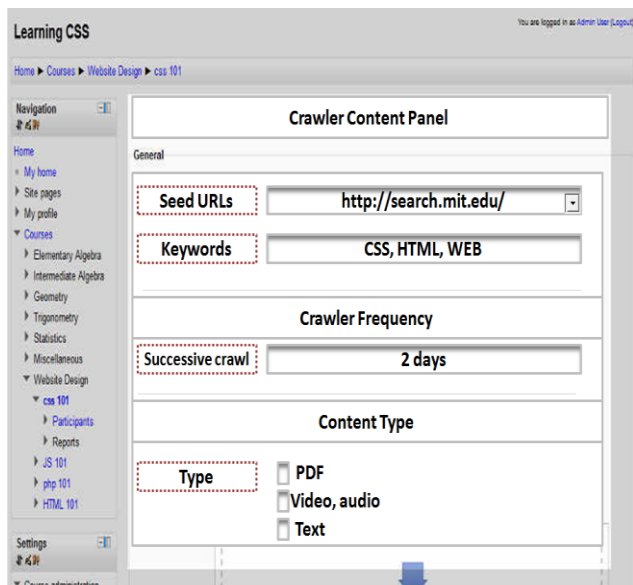


Fig. 3. Crawler content panel

Consequently, as shown in Figure 4, the crawler visits the source file of the specified URL (e.g., search.mit.edu), the recognizes the CSS as a keyword, and fetch the URLs under the meta tag “<a href = ...” as a potential content resource.



Fig. 4. Source file of search.mit.edu

B. The Stage of Teacher Recommendations

The teacher has the ability to upload digital resources of his own, colleagues, OERs ... associated to an eCourse into the Repository of Digital Resources. These Resources are recommended ones and mainly the best ones.

After the teacher uploads any resource, the system is deleting any duplicate. Also, the system gives the teacher resources high priorities.

C. The Stage of Filtering

When the students enter his eCourse, this stage will be activating. This Stage consists of:

- Collaborative Filtering: this approach will organize the priorities of the recommendations. The general mechanism based on defining subgroups (every subgroup known as the nearest neighbors) whose preferences are similar to the active user, so the nearest neighbors of the active student are those students who share the same institute (department, school). Then this stage calculates the average of the subgroups rating to order the recommendations upon the high rates.
- The Rating Matrix: The target CMS need to have a way to capture the rating by explicit, implicit methods or mixture of them. These students' rates of the digital resources saved in the CMS database as a table of two dimension matrix; where the row represents all the rates of one student on all digital resources while the column represents all the rates of all students on one digital resource (see table I).

Table I. rating matrix

<i>Student</i> \ <i>Digital Resources</i>	<i>Dr1</i>	<i>Dr2</i>	...	<i>Drn</i>
Std1				2
Std2	5	3	3	
...			3	5
Stdn	3		5	

- The general steps of Collaborative Filtering are:
 - a) Receiving the list of the recommended digital resources.
 - b) Finding the neighbors of the active student.
 - c) Finding the average rates of the neighbors for every recommend digital resource.
 - d) Organizing the recommendations upon the highest average; firstly, organizing the set of the “teacher recommendations” which already have the higher priorities then organizing the other recommendations set which came from Web.
 - e) Finally, the “recommended digital resources” are passed to the next steps.

- **Demographic-Based Filtering:** Theoretically, the role of this approach is to filter the incoming recommendations upon the students' demographic (and personal) data that related to education issues. For example, the following demographic-personal data could be related to the education issues: preferred language, student specialization, study level year, faculty, and department. The language filtration as an example means that the active student needs all the recommended digital resources in his preferred language, so any language of digital resources in the recommendations list defer from his preferred language will be deleted.
- The general steps of Demographic-based Filtering are:
 - a) Receiving the list of the recommended digital resources.
 - b) Reading the related demographic and personal data of the active student profile.
 - c) Matching the related fields of each digital resource from the list with the fields of the active student profile, so if the matching process is not positive; the digital resource will be deleted from the list.
 - d) Finally, the "recommended digital resources" are passed to the next steps.
- **Rule-Based Filtering:** It will filter the incoming recommended digital resources upon a set of rules, which could be found in the student profile and in the system profile. The system administrator put some rules in the system profile, while the student can put his own rules in his profile.

We suggest that the following types of rules that could be used in the student profile and the system profile to filter the listed digital resources (see figure 5):

- 1) **Link:** the system will filter out any digital resource whose link found in the rules profiles.

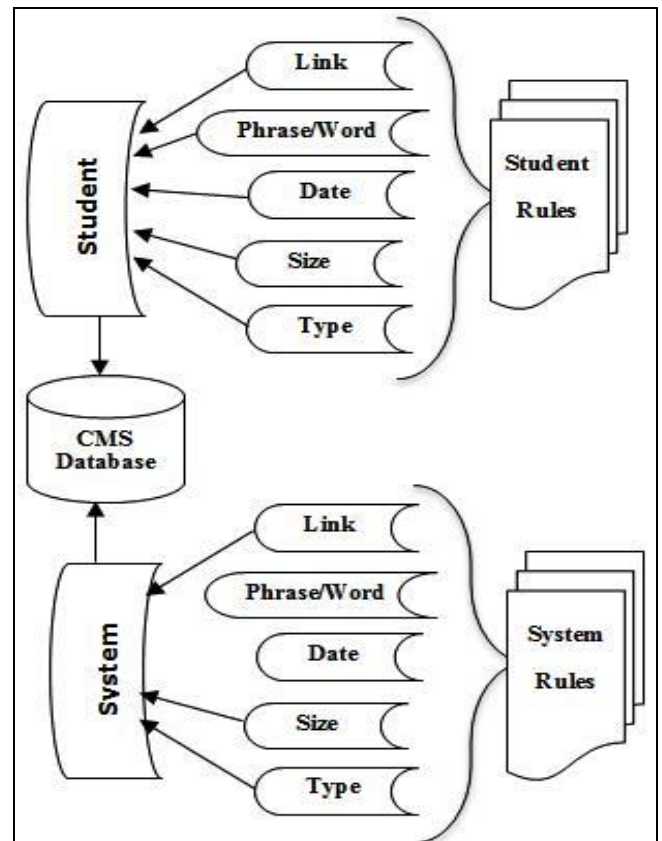


Fig. 5. Student and system rules

- 2) **Phrase or word:** the system will filter out any digital resource which his name, keywords or abstract match any "phrase" or "word" found in the rules profiles.
- 3) **Date:** the system will not show any digital resource does not fit the date criteria.
- 4) **Size:** the system will not show any digital resource does not fit the size criteria.
- 5) **Type:** the system will not show any digital resource does not fit the type criteria.

- The general steps of Rule-Based Filtering are:
 - a) Receiving the list of the recommended digital resources.
 - b) Reading the following fields of the system rules:
 - 1) Field which contains link of digital resource.
 - 2) Field which contains keywords.
 - 3) Fields of maximum and minimum dates.
 - 4) Field which contains the allowed size.
 - 5) Field which contains the forbidden types.

- c) The system deletes from the recommendations list every digital resource that matches any link or keywords as well as any digital resource whose dates are out of the minimum-maximum dates. It also deletes any digital resource, whose size is larger than the allowed size and whose type matches the forbidden types.
- d) Reading the same fields of rules from the student profile and repeating the filtration process.
- e) Finally, the “recommended digital resources” are prepared to be presented in a suitable way on the windows of active student eCourse.

V. DISPLAYING RESULT VIA CMS/LCMS

Displaying results to students could be done within many methods. As a case, it could be displayed via eLearning platform (LMS, CMS and LCMS). Respect to our paper, the appropriate platform is the open source one which allow adding new part within the course page to display the results when the tutor/student accessing his course. Fortunately there are many Open-Source LCMSs, e.g.

- MOODLE <http://moodle.org>
- ILIAS <www.ilias.de/ios/index-e.html>
- Claroline <www.claroline.net>

As an example, we could use MOODLE in our case.

The procedure could be the following:

- a) When the user (student or tutor) logging into his eCourse, Moodle retrieves the existing resources in the repository which related to the current eCourse.
- b) The RS activate the stage of Filtering and finishing with a list of recommended digital resources.
- c) The list is showing in a “block” in the eCourse page.

Figure 4 shows this example block within Moodle course page.

VI. CONCLUSION

RSs have been widely used in many Internet activities, mainly to overcome the information overload problem, which the user faced while searching any item and getting thousands of unrelated results. This research tries to solve the overload problem when students searches Web about suitable and related digital resources to their current eCourse as well as preferable resources to their needs and taste.

This research summarizes some essential information about crawler system, CMS/LCMS and RS. In addition, it reviews the suitability of RS approaches to recommend digital resources from Web to CMS/LCMS. Furthermore, the paper studies and presents a new RS algorithm to recommend suitable digital OERs from Web to students while entering an eCourse in CMS/LCMS. These proposed algorithm is considered as a Hybrid Recommender System which consist of some RS approaches; Content-Based System, Collaborative Filtering, Rule-Based Filtering and Demographic-Based Filtering

REFERENCES

- [1] Abel Caine. 2012. Global list of OER initiatives / Open Educational Resources (OER)'s Directories, WSIS – UNESCO. Created by Abel Caine 23 Nov. Retrieved January 28, 2015. www.wsis-community.org/pg/directory/owned/group:14358.
- [2] Andronico, A., et. 2003. Integrating a multi-agent recommendation system into a Mobile Learning Management System. *Artificial Intelligence in Mobile System 2003- AIMS 2003*, In conjunction with *UbiComp 2003*, October 12th-15th, Seattle, USA. Retrieved December 1, 2009 from http://w5.cs.uni-sb.de/~krueger/aims2003/camera-ready/carbonaro-4.pdf
- [3] Calvo, R. 2003. User Scenarios for the design and implementation of iLMS, In Proceedings of the AIED2003 workshop, Towards Intelligent Learning Management Systems, p.115-123, July 20. Retrieved December 1, 2009 from http://www.cs.usyd.edu.au/~aied/vol4/vol4_calvo.pdf
- [4] Chakrabarti, Soumen, Martin Van den Berg, and Byron Dom. 1999. "Focused crawling: a new approach to topic-specific Web resource discovery." *Computer Networks* 31.11 (1999): 1623-1640.
- [5] Francesco Ricci and Lior Rokach and Bracha Shapira. 2011. *Introduction to Recommender Systems Handbook*, Springer, pp. 1-35.
- [6] Geser, G. 2012. (ed.): *Open Educational Practices and Resources – OLCOS Roadmap* 2012, www.olcos.org/cms/upload/docs/olcos_roadmap.pdf . Retrieved January 28, 2015.
- [7] Itmazi, J. & Gea, M. 2006. *The Recommendation Systems: Types, Domains and the Ability Usage in Learning Management System*. the



Fig. 4. Block of updated resource in Moodle course [28]

- International Arab Conference on Information Technology (ACIT'2006). Yarmouk University, Jordan. Dec. 19th–21st.
- [8] Itmazi, J. & Gea, M. 2008. Using Recommendation Systems in Course Management Systems to recommend Learning Objects, *IAJIT*, Vol. 5, No. 3, July 2008, p234-240. ISSN: 1683-3198.
- [9] Itmazi, J., Gea, M., Paderewski, P. & Gutiérrez, F. 2005. A Comparison and Evaluation of Open Source Learning Management Systems. *IADIS International Conference-Applied Computing 2005*, Algarve, Portugal. February 22nd-25th. ISBN: 972-99353-6-X.
- [10] Jamil Itmazi. 2010. A suggested algorithm of recommender system to recommend learning objects from digital library to learning management system. *Asian Journal of Information Technology*. Volume: 9. Issue: 2. Page No.: 37-44. Year: 2010. DOI: 10.3923/ajit.2010.37.44 URL: <http://medwelljournals.com/abstract/?doi=ajit.2010.37.44>
- [11] Koren, Y., Bell, R.M., Volinsky, C. 2009.: Matrix factorization techniques for recommender systems. *IEEE Computer* 42 (8), 30–37.
- [12] Lu, J. 2004. A Personalized e-Learning Material recommender System. the 2nd International Conference on Information Technology for Application (ICITA 2004). HARBIN, CHINA: January 9th-11th. pp. 374-379. ICITA2004 ISBN 0-646-42313-4.
- [13] McIntosh, Don. 2015. Vendors of Learning Management and E-learning Products. Trimeritus eLearning Solutions Inc. Updated January 6, 2015. Retrieved January 28, 2015. www.trimeritus.com/vendors.pdf
- [14] MIT. 2014. Site Statistics, MIT OpenCourseWare. http://ocw.mit.edu/about/site-statistics/monthly-reports/MITOCW_DB_2014_03.pdf. Retrieved January 28, 2015.
- [15] Nageswara, R. & Talwar, V. 2008. Application Domain and Functional Classification of Recommender Systems—A Survey. *DESIDOC Journal of Library & Information Technology*, Vol. 28, No. 3, May 2008, pp. 17-35. ISSN: 0971-4383.
- [16] Qing Chen. 2010. Use of Open Educational Resources: Challenges and Strategies.. *ICHL*, volume 6248 of *Lecture Notes in Computer Science*, page 339-351. Springer.
- [17] Robin D. 2002. Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-Adapted Interaction*. Vol. 12, No. 4. (1 November 2002), pp. 331-370. ISSN:0924-1868, Springer. DOI: 10.1023/A:1021240730564
- [18] Sabic, A.; El-Zayat, M. 2010 . Building E-University Recommendation System, The 2nd IEEE International Conference on Information management and engineering (IEEE ICIME 2010). 978-1-4244-5263-7
- [19] Schafer J., Konstan J. & Riedl J. 2001. E-commerce recommendation applications. *Data Mining and Knowledge Discovery*, Vol. 5 , Issue 1-2, p115-153, 2001, Springer Netherlands Publishers, DOI: 10.1023/A:1009804230409.
- [20] Schmitz, Christoph, et al. 2002. "Accessing distributed learning repositories through a courseware watchdog." *Proc. of E-Learn 2002 World Conference on E-Learning in Corporate, Government, Healthcare, & Higher Education (E-Learn 2002)*.
- [21] Taghipour, N., Kardan, A.. 2008. A hybrid web recommender system based on q-learning. In: *Proceedings of the 2008 ACM Symposium on Applied Computing (SAC)*, Fortaleza, Ceara, Brazil, March 16-20, 2008, pp. 1164–1168.
- [22] Tane, Julien, Christoph Schmitz, and Gerd Stumme. 2004. "Semantic resource management for the web: an e-learning application." *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*. ACM.
- [23] Tane, Julien, et al. 2003. "The Courseware Watchdog: an Ontology-based tool for finding and organizing learning material." *Fachtagung Mobiles Lernen und Forschen* 6.11.
- [24] Terveen, L. & Hill, W. 2001. Beyond Recommender Systems: Helping People Help Each Other. In J. M. Carroll (Ed.) *Human-Computer Interaction in the New Millennium*. ISBN: 0201704471, pp. 487-509. New York. Addison-Wesley. ACM Press.
- [25] Vozalis, M. & Margaritis, K. 2004. Collaborative Filtering enhanced by demographic correlation. the *AIAI Symposium on Professional Practice in AI*, of the 18th World Computer Congress, Toulouse, France, August 22nd -27th. pp. 393-402. DOI: 10.1.1.95.8507.
- [26] Welsh, Elizabeth T., et al. 2003. "E-learning: emerging uses, empirical results and future directions." *International Journal of Training and Development* 7.4: 245-258.
- [27] Yang, Stephen JH. 2006. "Context aware ubiquitous learning environments for peer-to-peer collaborative learning." *JOURNAL OF EDUCATIONAL TECHNOLOGY AND SOCIETY* 9.1 (2006): 188.
- [28] W. Hijazi, Haytham, and Itmazi Jamil. 2013. Crawler Based Context Aware Model for Distributed e-Courses through Ubiquitous Computing at Higher Education Institutes. 4th International Conference on. IEEE.

Temperature Aware Design for High Performance Processors

Mustafa M. MUSTAFA, Muhammed A. IBRAHIM, Diary R. SULAIMAN

Electrical Engineering Department
 Engineering College, Salahaddin University-Hawler
 Erbil-Kurdistan-Iraq

e-mail: mmmmt87@gmail.com, mabdulbaki@hotmail.com, diary.r.sulaiman@su.edu.iq

Abstract—efficient temperature aware design in modern microprocessors, especially in the design of digital portable, notebook, and handheld computers is becoming increasingly important. Many studies have been done on microprocessor’s dynamic thermal management techniques and methodologies; from thermal estimation to voltage scaling, clock gating, and total/active power monitoring and control. As technology, moves into deep submicron feature sizes and the leakage power are expected to increase because of the exponential increase in leakage currents with technology scaling. In nanometer technologies, it is observed that leakage power will become comparable to dynamic or total power dissipation in the next generation processors in the next few years. This paper presents a hardware design for dynamic thermal management strategies for microprocessors leakage power control, which is particularly appealing for portable and embedded systems. LTspice simulation program is used to verify the theoretical idea and confirm the design operations. Results shows that, the appropriate thermal management system can be designed for a much lower maximum power rating with minimal performance impact for typical applications, considerable amount of power consumption reduction as well as thermal aware challenges have been obtained.

Keywords— Thermal Management, Leakage Power, Reverse Body Biasing.

I. INTRODUCTION

Through past two decades, microprocessor performance has matured about one thousand, delivering extraordinary computing capabilities[1]. The evolution in microprocessor performance has predominantly be driven by the regular scaling of the transistor process featuring that expedites capacity of exponential transistor integration. In other words, the number of transistors in a dense integrated circuit doubles approximately every two years according to Moor’s law. Increase in device count per chip and shrinkage of feature size are shown in Fig. 1 [2,3]. More than one-million fold increase in the device count has been achieved these 40 years leading to almost the same increase in processor performance. Whenever transistor density increases, the power density in the microprocessors has also extended exponentially [3]. Further, this power, which is consumed by a microprocessor or integrated circuits, is dissipated as heat due to the resistive behaviour of the Complementary Metal Oxide Semiconductor (CMOS) circuits. High temperature has an adverse effect on many perspectives of microprocessors, such as transistor performance, power consumption and system reliability.

Whenever transistor density increases, the power density in the microprocessors has also extended exponentially[3]. Further, this power, which is consumed by a microprocessor or integrated circuits, is dissipated as heat due to the resistive behaviour of the Complementary Metal Oxide Semiconductor (CMOS) circuits. High temperature has an adverse effect on many perspective of

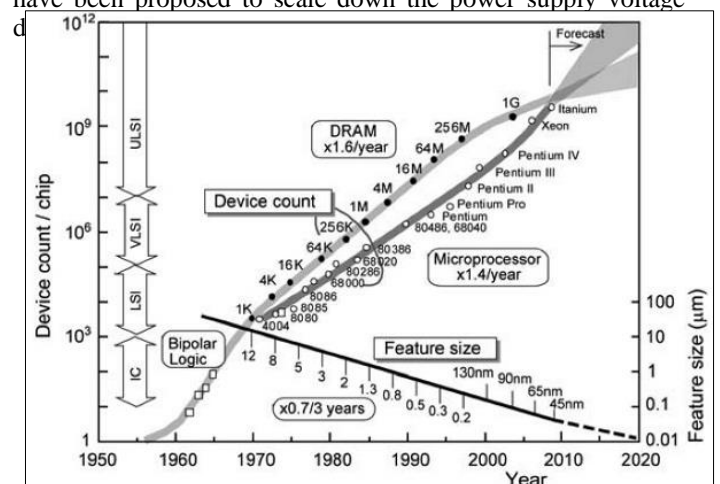
microprocessors, such as transistor performance, power consumption and system reliability.

There are Two main components of power dissipation in CMOS circuits, as described in equation (1) [4].

$$P = P_{dynamic} + P_{leakage} \quad (1)$$

Where P is the total power dissipation, $P_{dynamic}$ is the dynamic power, and $P_{leakage}$ is the static/leakage power.

Lowering power supply voltage is one of the effective schemes to reduce the power dissipation. A number of methods have been proposed to scale down the power supply voltage



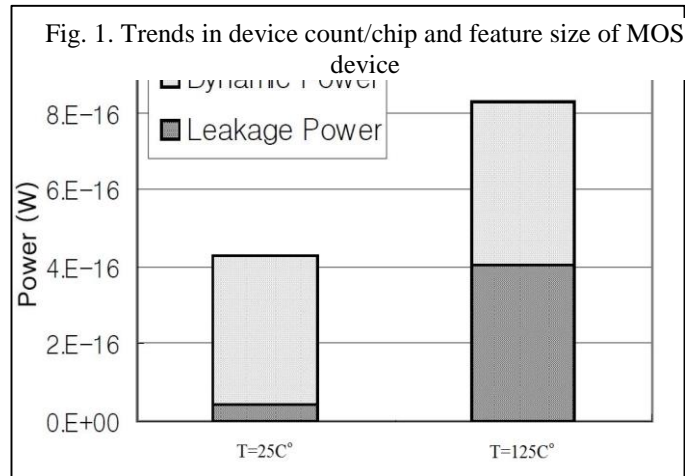


Fig. 2. Dynamic and leakage power of CMOS inverter for different temperatures using the 70nm technology

Even though they are effective in decreasing dynamic power dissipation, it does not help reducing the leakage power effectively, which is the most responsible source for generating heat in CMOS circuits. The non-ideal off-state/leakage characteristic, or a finite resistance, of MOSFETs, current drawn from a power supply even when a transistor operates in the cutoff region. This makes microprocessor and CMOS circuits to consume a power even when they are in standby mode, which is a leakage/static power. Fig. 2 shows the dynamic and leakage power of a 70nm inverter for different operating temperatures. The leakage power, which was initially 10% of the total power at room temperature, increases up to 49% as the temperature goes up to 125C°[6].

As transistor geometries are shrunk aggressively, threshold voltage decreases to achieve high performance resulting in exponential increase in leakage current. Due to the continued scaling of technology, supply and threshold voltage, and leakage, power has become dominant in the power dissipation of nanoscale CMOS circuits. Therefore, optimal power dissipation for a given performance depends not only on power supply voltages but also on the device threshold voltage [7].

To effectively limit the high temperature and reduce the leakage power with a cost-effective, threshold voltage has to be increased in circuit level, by designing a dynamic hardware controller. Body biasing is a low power/thermal management technique used in microprocessors and CMOS digital circuits. It offers an optimum solution for the CPU power/thermal problems by CMOS transistors during modifying circuit operation or program execution. This work uses a reverse body biasing technique for controlling the temperature of microprocessor.

The models and techniques of minimizing dissipated power and generated temperature/thermal variation of a high performance modern CMOS processor are underway, and it has

prompted a new research area and in low-power in conjunction with low temperature techniques. Such as, analysis of thermal dynamics in multicore systems, focusing on time response to heat generated in a single core and its connection with very large scale integration (VLSI) design principles that's presented in [8]. Qikai Chen in [9] stabilized the chip temperature by online monitoring using temperature sensor. The proposed circuits in [10] and [11] adaptively adjust the body bias to its optimal value during the whole standby period, which result in considerable reduction of leakage power and effective compensation of process and temperature variations. . Nikhil Saxena and Sonal Soni in [23] present that circuit level techniques incorporated requiring support from technology and process level techniques can be more effective in reducing leakage, where they used stacking effect to reduce leakage current for different input vectors for a stack of 3 Nano technology NMOS transistors. Designs of Forward Body Biased Adiabatic Logic for reduction of average, peak, and differential power have been presented in [24]; this body biasing method was applied to adiabatic Toffoli and Fredkin gates. The designs improved over their non-body biased implementation in all power metrics, as well as improved output signals. The designs improved the differential power over conventional NAND and MUX gates.

In order to ensuring low power operation for efficient power and temperature aware design for modern microprocessors, this work tries to study, design, and simulate the temperature aware controller based on dynamic frequency scaling. Where this controller adjust the threshold voltage of NMOS transistors within the microprocessor by sensing it's temperature, then generating frequencies matching the frequency of the microprocessor to make the controller respond to the change in temperature with a quickness proportions of the microprocessor's speed. Resulted frequency will actuate the controller to produce a voltage, which is reverse body bias voltage (V_{body}).

II. EFFECT OF BODY BIASING ON THRESHOLD VOLTAGE, LEAKAGE COMPONENTS AND TEMPERATURE

The threshold voltage is typically adjusted and accommodated during the fabrication process by varying the doping concentration in the channel area [12]. Alternatively, the body bias circuit technique utilizes the body terminal to dynamically modify the threshold voltage of a transistor during circuit operation. Depending upon the polarity of the voltage difference between the source and body terminals (VSB), the threshold voltage can be either increased or decreased as compared to a zero body biased transistor. The threshold voltage is increased when the source-to-substrate p-n junction of a MOSFET is reverse biased. The threshold voltage of a MOSFET can also be reduced by forward biasing the source-to substrate p-n junction [13]. Among the leakage power reduction techniques, RBB technique, which increases the threshold voltage (V_{th}) of transistors and are extensively employed to suppress the sub-threshold leakage current (ISUB) as depicted in Figs. 3 and 4 [14,15].

However, this technique also aggravates short channel effects (SCEs), such as drain-induced barrier lowering (DIBL), gate-induced drain leakage (GIDL), and band-to-band tunneling (BTBT) current. In particular, GIDL and BTBT current significantly increases under the reverse body bias condition since the state-of-the-art MOSFETs are fabricated with high overall doping concentrations, lowered source/drain junction depths, halo doping, high-mobility channel materials, etc.

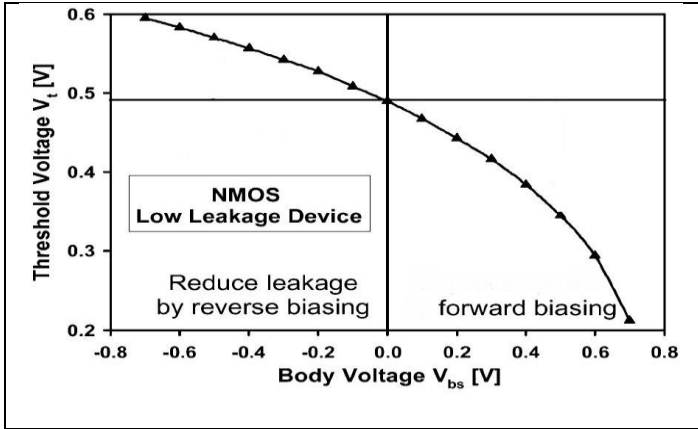


Fig. 3. Body bias effect on threshold voltage in 90-nm CMOS technology

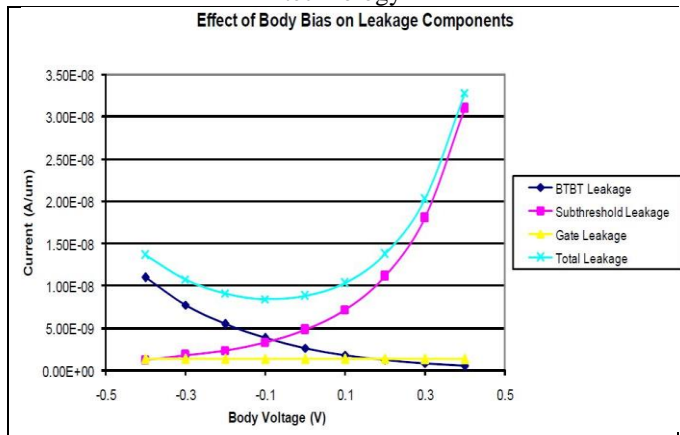


Fig. 4. Effect of body bias on leakage components for a 70nm predictive technology

Furthermore, the reduction of the gate oxide thickness (t_{ox}) causes a drastic increase in the gate tunnelling leakage current due to carriers tunnelling through the gate oxide, which is a strong exponential function of the voltage magnitude across the gate oxide [16,7].

When the gate-to-source voltage in a MOS transistor is below V_{th} , the transistor is not completely turned off, instead, there is weak inversion region having some minority carriers along the length of the channel.

This makes a small current flow from drain to source in NMOS case, which is called the sub-threshold leakage. I_{SUB} is the component that affected by temperature and threshold voltage. This current can be expressed based on [14]:

$$I_{SUB} = \mu C_{dep} \frac{W}{L} V_T^2 \left(\exp \frac{V_{GS} - V_{th}}{nV_T} \right) \left(1 - \exp \frac{-V_{DS}}{V_T} \right) \quad (2)$$

Where $C_{dep} = \sqrt{\epsilon_{si} q N_{sub} / (4\phi_B)}$ denotes the capacitance of the depletion region under the gate area, ϵ_{si} is the permittivity of Si, q is the electron charge, N_{sub} is the doping concentration of the p-substrate, ϕ_B is the built-in potential, V_T is the thermal voltage that is equal to kT/q , C_{ox} is the oxide capacitance per unit area between the gate metal and the bulk surface, and n is the sub-threshold parameter and is expressed as $1 + C_{dep}/C_{ox}$.

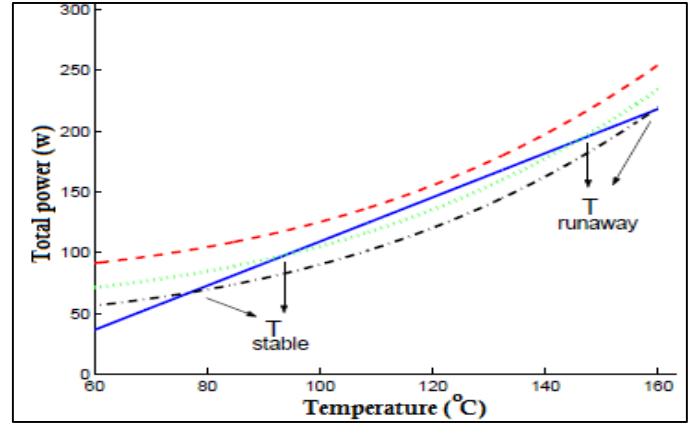


Fig. 5. The power-temperature thermal runaway

From the above equation that the positive feedback loop between power dissipation and temperature is seen i.e., increasing the temperature will increase the leakage current and increased leakage current will further increase the temperature. Higher generated temperatures lead to higher dissipated leakage and total power, therefore thermal runaway will occur when the rate of temperature generation becomes greater than the rate of heat removal [15].

Power-temperature thermal runaway will become more critical in future technologies.

Fig. 5 shows that, the thermal runaway is a destructive positive feedback condition that can occur when inadequate thermal control is combined with a silicon process technology where the leakage increases the current exponentially with the generated temperature.

III. CONTROLLER IMPLEMENTATION

The specific controller implementation of the proposed technique is shown in Fig. 6. The temperature of the microprocessor chip will be directly converted to a frequency, which match the frequency of the desired microprocessor or CMOS chip. After that, the phase locked loop (PLL) will process this frequency to generate the RBB voltage.

In microprocessors, the non-uniformity of power consumption can cause a much higher local power density. Those regions on the die with a high temperature called hotspots. Temperature in the hotspots rise much faster than the full chip; therefore, they could be utilized for temperature sensing by using temperature sensors [16]. A voltage signal will be

generated by temperature monitoring circuit and the value of voltage signal will increase proportionally with temperature [9].

To make the controller system operate with accurate response, the incoming voltage from temperature sensor has to be converted to a higher frequency, matching the frequency of the desired microprocessor. To achieve this purpose, voltage control oscillator (VCO) could be used [17]. After that the temperature is converted to frequency by temperature monitoring circuit, VCO and frequency divider, PLL shown in Fig. 6. will be driven by this frequency to generate a voltage, which is the voltage at the output of the loop filter within PLL.

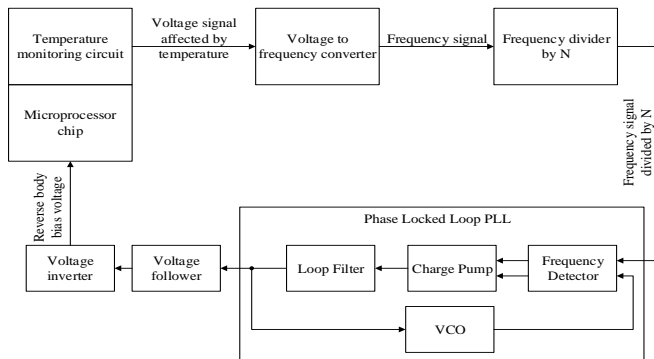


Fig. 6. Block diagram of temperature controller unit

This voltage will be the desired reverse body bias voltage that has been used to decrease leakage power/temperature of microprocessor or CMOS chip.

IV. SIMULATION RESULTS

Based on specification of microprocessor and its temperature usage levels, the controller computes the required V_{body} to be applied on the bodies of NMOS transistors within CPU. This means that, the temperature control unit is the power controller such that leakage power consumption and generated heat at a particular operating frequency are minimized. Table 1 shows the temperature usage levels of Intel® Core™ i3-3217U 1.80 GHz Processor (which is one of the modern microprocessor produced in recent years) [18].

By applying the controller on the mentioned CPU, the temperature control unit will be activated just when the temperature reaches maximum range, otherwise the controller will be in idle state. Jing Yang and Yong-Bin Kim approve that the optimal value of RBB is 0.38 [19], therefore in this work we will consider this value as a maximum value of RBB voltage generated by temperature controller unit. The maximum temperature range will be divided to three ranges to generate three values as presented in table 2.

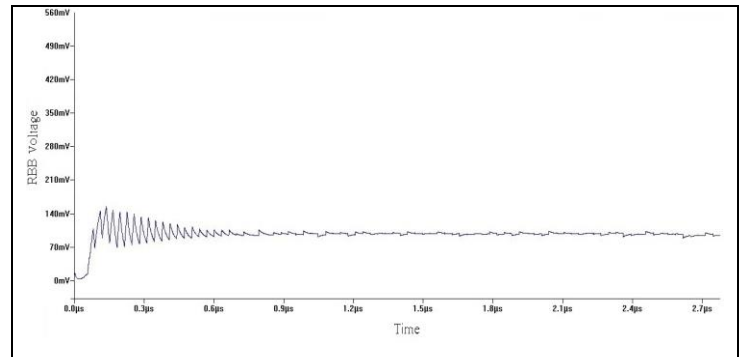
Table 2 confirms that the generated frequency identify the frequency of the desired CPU. Fig. 7. shows the simulation results of the controller for the three cases.

Table 1 Intel® Core™ i3-3217U temperature levels

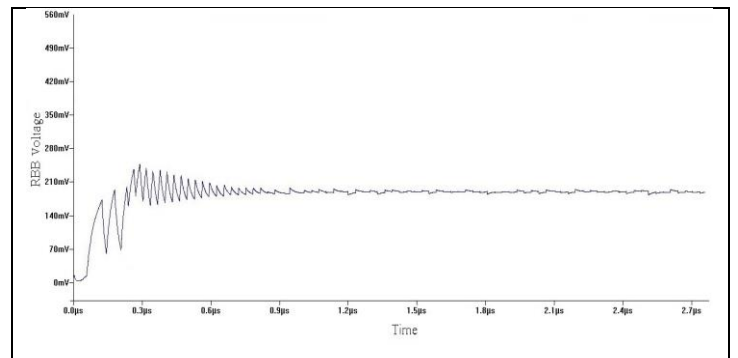
Idle temperature	Normal temperature	Maximum temperature
34 to 49 °C	50 to 60 °C	61 to 75 °C

Table 2 Desired value of temperatures for the controller

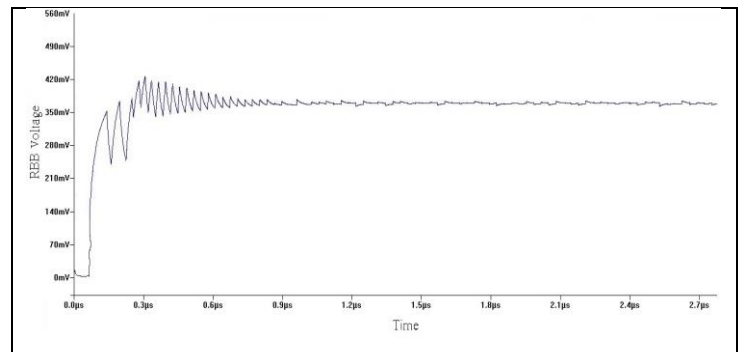
Temperature range	VCO frequency	RBB voltage
61 to 65 °C	1.6 GHZ	0.1 V
66 to 70 °C	1.7 GHZ	0.2 V
71 to 75 °C	1.8 GHZ	0.38 V



(a)



(b)



(c)

Fig. 7. RBB voltage with time for: (a) 61 to 65 °C; (b) 66 to 70 °C; (c) 71 to 75 °C;

By applying, those values of RBB voltage on a single CMOS inverter, leakage current and temperature will be decreased significantly. Fig. 8. shows the effect of RBB voltage on leakage current for different temperature ranges.

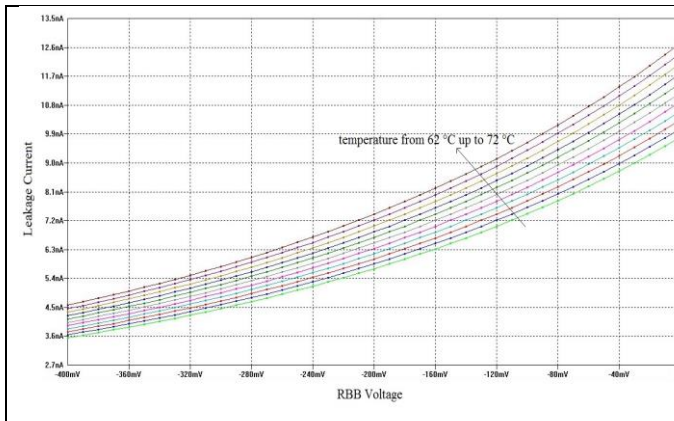


Fig. 8. Effect of RBB voltage on leakage current of a CMOS inverter using 50nm technology I for different

V. CONCLUSION

Thermal management is one of the many crucial tasks in the design of high performance processors/CMOS chips. Current processor's design challenges in portable and embedded systems becoming more complex and high transistor's count, both thermal management and power consumption reduction is becoming significant in the design process. Therefore, improving microprocessors power and temperature have been the primary consideration in digital VLSI design; many researches are now underway on minimizing power dissipation and operating temperature reduction, especially for portable and high performance systems. In portable systems where processor circuits spend most of their time in an idle state with no computation, to maintain both the processor's supply voltage and clock frequency, thermal reduction technique has emerged as an effective way for minimizing energy consumption as well as low temperature designs. In this study, a temperature control unit was designed and simulated using LTspice simulator for different temperatures' and it was confirm that, the controller generates the desired values of RBB voltage for different temperature ranges; those values were applied for a single CMOS inverter, using 50nm technology.

It was clear that the controller decreased the leakage current significantly, and because of the direct proportionality between leakage current and temperature, the temperature was automatically detracted.

Therefore, the temperature aware technique confirms superiority on leakage power consumption reduction as well as temperature reduction mechanisms, and it shows the good promises in processor's thermal-energy optimization. Therefore,

it is possible to integrate and use the thoughts of this technique with any existing processor to study system level power and thermal issues for various high performance processors accurately.

REFERENCES

- [1] S. Borkar and A. A. Chien, "The Future of Microprocessors," pp. 67–77, 2011.
- [2] K. Skadron, M. R. Stan, W. Haung, S. Velusamy, K. Sankaranarayanan, and D. Tarjan, "TEMPERATURE-AWARE COMPUTER SYSTEMS: OPPORTUNITIES AND CHALLENGES," IEEE Comput. Soc., vol. 23, no. 6, pp. 52–61, 2003.
- [3] S. Borkar, "Thousand Core Chips — A Technology Perspective," in Design Automation Conference, 2007, pp. 746–749.
- [4] P. F. Butzen and R. P. Ribas, "Leakage Current in Sub-Micrometer CMOS Gates."
- [5] T. Kuroda and Kojiro Suzuki, "Variable Supply Voltage Scheme for Low-Power High Speed CMOS Digital Design," IEEE J. Solid-State Circuits, vol. 33, no. 3, p. 1998.
- [6] C. H. Kim, K. Roy, and W. Lafayette, "Dynamic V_{TH} Scaling Scheme for Active Leakage Power Reduction," in Automation and Test in Europe Conference and Exhibition, 2002, pp. 2–6.
- [7] K. Roy, S. Mukhopadhyay, and S. Member, "Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits," vol. 91, no. 2, 2003.
- [8] S. Mikula, "Thermal dynamics of multicore integrated systems," in IEEE Transactions on Components and Packaging Technologies, 2010, pp. 524–534.
- [9] M. Meterelliyoz and K. Roy, "A CMOS Thermal Sensor and Its Applications in Temperature Adaptive Design," 7th Int. Symp. Qual. Electron. Des., pp. 243–248, 2006.

- [10] K. Xiao, C. Liu, and Y. Sun, "A Novel Adaptive Reverse body bias Technique to Minimize Standby Leakage Power and Compensate Process and Temperature Variation," in *ieee*, 2011, p. 15651568.
- [11] K. K. Kim and Y. Kim, "A Novel Adaptive Design Methodology for Minimum Leakage Power Considering PVT Variations on Nanoscale VLSI Systems," *IEEE Trans. Very Large Scale Integr. Syst.*, vol. 17, no. 4, pp. 517–528, 2009.
- [12] V. Kursun, *Supply and Threshold Voltage Scaling Techniques in CMOS Circuits*. 2004, p. 363.
- [13] A. Srivastava and C. Zhang, "An Adaptive Body-Bias Generator for Low Voltage CMOS VLSI Circuits," *Int. J. Distrib. Sens. Networks*, vol. 4, no. 2, pp. 213–222, 2008.
- [14] K. Von Arnim, E. Borinski, P. Seegebrecht, H. Fiedler, R. Brederlow, R. Thewes, J. Berthold, and C. Pacha, "Efficiency of Body Biasing in 90-nm CMOS for Low-Power Digital Circuits," *IEEE SOLID STATE CIRCUITS*, vol. 40, no. 7, pp. 1549–1556, 2012.
- [15] C. Neau and K. Roy, "Optimal body bias selection for leakage improvement and process compensation over different technology generations," *Proc. 2003 Int. Symp. Low power Electron. Des. - ISLPED '03*, p. 116, 2003.
- [16] H. Jeon, Y. Kim, and M. Choi, "Standby Leakage Power Reduction Technique for Nanoscale CMOS VLSI Systems," *IEEE Trans. Instrum. Meas.*, vol. 59, no. 5, pp. 1127–1133, May 2010.
- [17] S. A. Parke, S. Member, and J. E. Moon, "Design for Suppression of Gate-Induced Drain Leakage in LDD MOSFET's Using a Quasi-Two-Dimensional Analytical Model," vol. 39, no. 7, 1992.
- [18] F. Zanini, D. Atienza, C. N. Jones, and G. De Micheli, "Temperature sensor placement in thermal management systems for MPSoCs," in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, 2010, pp. 1065–1068.
- [19] W. Liu, "Power and Thermal Management of System on chip," *Technical University of Denmark*, 2011.
- [20] Y. A. Eken and J. P. Uyemura, "A 5.9-GHz Voltage-Controlled Ring Oscillator in 0.18- m CMOS," *IEEE J. Solid-State Circuits*, vol. 39, no. 1, pp. 230–233, 2004.
- [21] intel, "Intel Company." [Online]. Available: <http://www.intel.com/content/www/us/en/homepage.html>.
- [22] J. Yang and Y. Kim, "Self Adaptive Body Biasing Scheme for Leakage Power Reduction under 32nm CMOS Regime," *Int. J. Adv. Comput. Sci.*, vol. 3, no. 9, pp. 453–459, 2013.
- [23] N. Saxena and S. Soni, "Leakage current reduction in CMOS circuits using stacking effect," *Int. J. Appl. or Innov. Eng. Manag.*, vol. 2, no. 11, pp. 213–216, 2013.
- [24] M. Morrison and N. Ranganathan, "Forward body biased adiabatic logic for peak and average power reduction in 22nm CMOS," *Proceedings of the IEEE International Conference on VLSI Design*, 2014.

Image Watermarking using DC Component of DCT

Rifat Kurban
Dept. of Computer Engineering
Erciyes University
Kayseri, Turkey
rkurban@erciyes.edu.tr

Hakki Bozpolat
Informatics and Inf. Security Research Center
TUBITAK
Gebze, Turkey
hakki.bozpolat@tubitak.gov.tr

Florenc Skuka
Dept. of Computer Engineering
Erciyes University
Kayseri, Turkey
skuka.f@gmail.com

Abstract— In this paper, a robust watermarking method based on modifying the DC component of the DCT coefficient is presented. A random generated key sequence is embedded into the DC coefficients of 8x8 blocks with a scaling factor. Effect of the scaling factor under several attacks is evaluated. Experiments show that the method is robust without losing the transparency.

Keywords— Image watermarking, Robust watermarking, DCT, DC Component

I. INTRODUCTION

The acquisition and processing of digital information has become widespread in the recent years as a result of the increase in the use of internet both in terms of the number of users and the link speed. Consequently, the need has emerged for the protection of digital pictures against copying. One of the main methods which have been proposed for the protection of copyright is the watermarking of digital images [1].

Digital watermarking is defined as the hiding of digital data inside another digital data as a seal. Digital watermarking methods may be used for many different purposes including the protection of copyrights, data confirmation and data hiding [2]. Two different kinds of watermarking operations are implemented on digital information namely visible and invisible. In visible watermarking, the watermark is placed on an area of digital image in a way to be detected visually. In invisible watermarking, the watermark is hidden and can only be extracted by the person who embeds it with a definite algorithm [3].

Watermarking techniques are categorized into two as spatial and transform domain techniques according to the domain in which the watermark will be embedded. Spatial domain techniques embed the watermark by updating the pixel values of the source image [4-5]. Transform domain techniques embed the watermark in the source image by means of updating the coefficients of the transformation domains such as discrete Fourier transformation (DFT), singular value decomposition (SVD), discrete wavelet transform (DWT) and discrete cosine transform (DCT) [6-7]. In general, transformation domain methods are more resistant against the attacks in comparison to spatial domain methods [8]. Numerous watermarking techniques have been proposed for images, videos and particularly the subjects of the invisibility of the watermark and its robustness against the attacks of the enemy have been considered important [9]. In robust watermarking algorithms, it is aimed that the

information within the image to resist various image processing attacks and to be recognizable when it is recovered.

In [10], it has been proposed that the watermark becomes resistant when it is embedded in DC components in comparison to being embedded in AC components. Two pre-defined α scaling techniques have been used and evaluated accordingly to the block texture content.

In this study, the method recommended by Huang et. al. [10] is examined in detail. The effect of the scaling factor is expressed numerically and the robustness and transparency is analyzed by evaluating the DCT based watermarking technique, which modifies DC components, on different scaling factors and images.

II. DCT BASED WATERMARKING USING DC COMPONENT

DCT is a sinusoidal transform and similar to the Fourier transform however it only uses cosine terms and for that reason does not include complex components [11].

DCT is successfully used in many coding systems since it ensures the best energy distribution in the field of frequency [12].

A two-dimensional DCT is defined as follows [13]:

$$F(u, v) = p(u)p(v) \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} I(x, y) \cos\left(\frac{(2x+1)u\pi}{2M}\right) \cos\left(\frac{(2y+1)v\pi}{2N}\right) \quad (1)$$

where I is the input image, M and N are the image height and width, respectively, $p(u)$ and $p(v)$ are calculated as follows:

$$p(k) = \begin{cases} \sqrt{\frac{1}{N}} & \text{if } k = 0 \\ \sqrt{\frac{2}{N}} & \text{else} \end{cases} \quad (2)$$

In DCT, $F(0,0)$ is defined as DC (low frequency) component and contains a dominant value in comparison to other coefficients. Other components are defined as AC (high frequency) components. The classification of the components in a 8x8 block is indicated in the Figure 1.

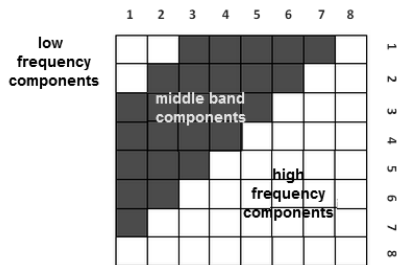


Fig. 1. DCT coefficients.

The flowchart of DTC based robust watermarking method is shown Figure 2.

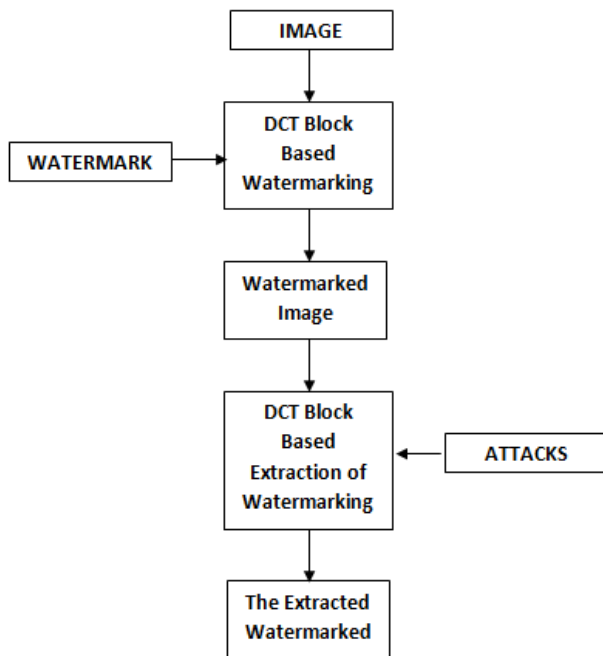


Fig. 2. DCT based watermarking: embedding and extraction.

A. Watermark Embedding

1. The original image I , is decomposed into blocks B of 8x8.
2. DCT is applied to all blocks individually.

3. The watermark w is created at random with a standard normal distribution as $\mu=0$ and $\sigma=1$

$$w = \{x_i, 0 \leq i \leq k\} \quad (3)$$

In equation (3), k is the number of blocks. The process of watermarking is realized only by updating the DC coefficients of each block. The embedding equation is as follows:

$$F'_k(u, v) = \begin{cases} F_k(u, v) \cdot (1 + \alpha \cdot x_k) & \text{if } u = v = 0 \\ F_k(u, v) & \text{otherwise} \end{cases} \quad (4)$$

In equation (4), α is the scaling factor.

4. Inverse discrete cosine transform (IDCT) has been applied on the blocks and the watermarked image WI has been obtained.

B. Watermark Extraction

1. The watermarked image WI and the original image I have been decomposed into blocks of 8x8.
2. DCT is applied to all blocks individually.
3. The differences of DC components in each block of the original image and watermarked image are calculated as the watermark w .

$$W_k^* = F_k^*(0,0) - F_k(0,0) \quad (5)$$

$$W^* = \bigcup_i W_k^* \{x_i^*, 0 \leq i \leq n\}$$

III. EXPERIMENTAL RESULTS

In robust watermarking it's aimed that to keep the level of degradation of the watermarked image as low as possible in comparison to the original picture (transparency) and to ensure that the watermark which has been extracted from the watermarked image as a result of attacks to resemble the original watermark as much as possible (robustness).

In this study, numerous experiments have been carried out using two grey level images with dimensions of 256 x 256 as shown in Figure 3. In the first case, host images are watermarked using different α values. It has been analyzed that to what extent the watermarked images have been degraded and to what extent the extracted images resembles to the original watermark. In another case, various attacks have been applied on watermarked images by using different α values and the robustness and transparency was analyzed.

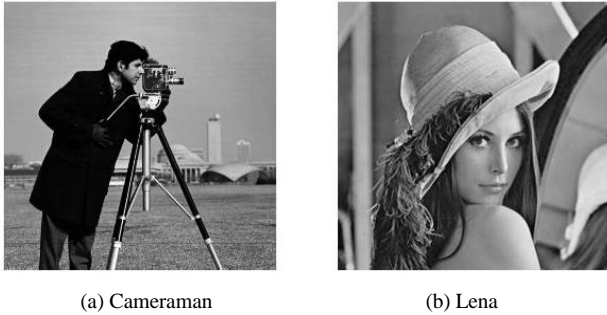


Fig. 3. Original host images for watermarking experiments.

A. Effect of the α parameter

In the experiments implemented in this section, the effect of different α values between 0.005 and 0.030 has been analyzed. Host images that have been watermarked by different α values are shown in Figure 4. Only the results for $\alpha = 0.005, 0.015$ and 0.030 are shown due to the lack of space. When the visual results are evaluated, it has been seen that unwanted blocking effect arises in the watermarked images as the α value increases.

In Table 1, numerical results of correlation analysis are given to show that the similarity of the original watermark and the extracted watermark As seen from the Table 1, the degree of similarity of the extracted watermark to the original watermark increases as the scaling factor (α) increases.

In Table 2, transparency of the original host image and the watermarked image has been analyzed by means of PSNR values. Choosing a high scaling factor (α) causes degradation in the image.



Fig. 4. Images watermarked with different α values.

TABLE I. ROBUSTNESS ANALYSIS FOR CAMERAMAN AND LENA ACCORDING TO SCALING FACTOR (α).

CORR	Cameraman	Lena
$\alpha=0.005$	0.8206	0.8368
$\alpha=0.010$	0.8857	0.9232
$\alpha=0.015$	0.8819	0.9364
$\alpha=0.020$	0.8936	0.9427
$\alpha=0.025$	0.9004	0.9390
$\alpha=0.030$	0.9037	0.9464

B. Effect of Attacks and the α parameter

In this section, four different attacks have been carried out, namely blurring(BL), JPEG compression(JC), salt & pepper noise(SP) and re-scaling(RS), for the purpose of testing the resistance of the watermarking method against attacks. Visual results and numerical analysis have been presented.

TABLE II. TRANSPARENCY ANALYSIS FOR CAMERAMAN AND LENA ACCORDING TO VARIOUS SCALING FACTORS (α).

PSNR	Cameraman	Lena
$\alpha=0.005$	117.2014	117.0920
$\alpha=0.010$	104.3893	104.3280
$\alpha=0.015$	95.9988	96.1121
$\alpha=0.020$	91.2254	90.5543
$\alpha=0.025$	87.1736	86.5733
$\alpha=0.030$	83.6518	82.3390

Due to limited space, the attacks that have been applied to the watermarked image only for the $\alpha = 0.015$ have been given for *Cameraman* and *Lena*, respectively, in the Figure 5 and 6.



(a) Blurring

(b) Jpeg Compression

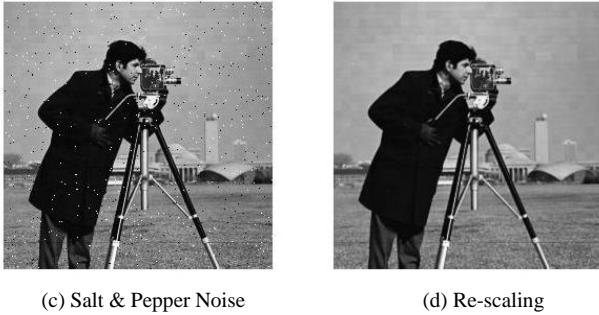


Fig. 5. Attacks on the watermarked images while $\alpha = 0,015$ for *Cameraman* image.

In Table 3, the similarity of the watermarks, which have been extracted after various attacks, to the original watermark has been obtained by correlation values. For both of the test images, correlation values have been obtained as higher than 0.25 even at the lowest value of $\alpha = 0.005$ and it was determined that the watermark is present in the image. On the other hand, the correlation values were over 0.8 and the existence of the watermark was undisputable at the value 0.030.

IV. CONCLUSION

In this study, effect of scaling parameter (α) on watermark extraction and degradation of image in robust watermarking of images has been examined by means of the modification of DC components in DCT. The effect of the watermarking method and α parameter has been analyzed by applying 4 different attacks. The results obtained have indicated that an appropriate α value must be chosen in order to ensure that the extracted watermark has a high degree of robustness and transparency.

ACKNOWLEDGEMENTS

The authors would like to thank Research Foundation of the Erciyes University, Kayseri, Turkey for supporting this work under the Grant No. FYL-2015-5826.

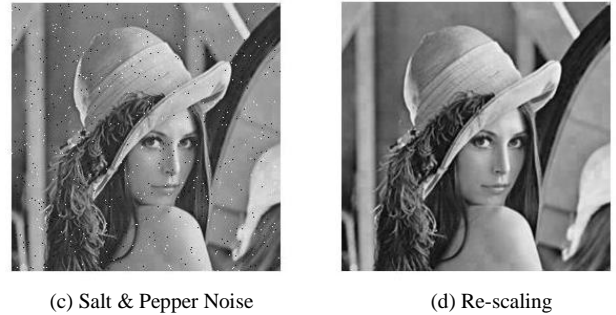


Fig. 6. Attacks on the watermarked images while $\alpha = 0,015$ for *Lena* image.

TABLE III. ROBUSTNESS ANALYSIS AGAINST VARIOUS ATTACKS FOR CAMERAMAN AND LENA ACCORDING TO THE SCALING FACTOR

CORR	Attacks			
<i>Cameraman</i>	<i>BL</i>	<i>JC</i>	<i>SP</i>	<i>RS</i>
$\alpha=0,005$	0.2857	0.6545	0.2735	0.7949
$\alpha=0,010$	0.5479	0.8045	0.5474	0.8809
$\alpha=0,015$	0.6423	0.8475	0.6757	0.8779
$\alpha=0,020$	0.7217	0.8733	0.7190	0.8914
$\alpha=0,025$	0.7739	0.8864	0.7646	0.8978
$\alpha=0,030$	0.8127	0.8906	0.8231	0.9031
<i>Lena</i>	<i>BL</i>	<i>JC</i>	<i>SP</i>	<i>RS</i>
$\alpha=0,005$	0.3443	0.6549	0.3164	0.8337
$\alpha=0,010$	0.5963	0.8405	0.5487	0.9231
$\alpha=0,015$	0.7213	0.9038	0.7276	0.9358
$\alpha=0,020$	0.8118	0.9174	0.8062	0.9424
$\alpha=0,025$	0.8338	0.9257	0.8264	0.9385
$\alpha=0,030$	0.8773	0.9390	0.8657	0.9466

REFERENCES

- [1] K. Prachi, L. Siddharth and T. Shreya, "Digital Watermarking for Protection of Intellectual Property", *IJCEM International Journal of Computational Engineering & Management*, Vol. 12, April 2011, ISSN (Online): 2230-7893.
- [2] G.C. Langelaar, I. Setyawan and R.L. Lagendijk, "Watermarking digital image and video data", *IEEE Signal Processing Magazine*, pp. 20-43, September 2000.
- [3] C.S. Shieh, H.C. Huang, F.H. Wang, and J.S. Pan, "Genetic watermarking based on transform domain techniques", *Pattern Recognition*, vol. 37, pp. 555-565, 2004.
- [4] Y.H. Yu, C.C. Chang and Y.C. Hu, "Hiding secret data in images via predictive coding", *Pattern Recognition*, vol. 38, pp. 691-705, 2005.
- [5] C.C. Chang, T.S. Chen and L.Z. Chung, "A steganographic method based upon JPEG and quantization table modification", *Information Sciences*, vol. 141, pp. 123-38, 2002.
- [6] C.T. Hsu and J.L. Wu, "Multiresolution watermarking for digital images", *IEEE Trans Circuits Systems II: Analog Digital Signal Process*, vol. 45, pp. 1097-101, 1998.
- [7] G.J. Yu, C.S. Lu and A. Mihy, "Message-based cocktail watermarking system", *Pattern Recognition*, vol. 36, pp. 957-68, 2003.
- [8] V. Aslantas, "A singular value decomposition based image watermarking using genetic algorithm", *Int J Electron Commun (AEU)*, 2007.
- [9] Y. Xueyi, C. Xueting, D. Meng, H. Shuyun, W. Yunlu, "A Multiple-Level DCT Based Robust DWT-SVD Watermark Method," *Computational Intelligence and Security (CIS)*, 2014 Tenth International Conference on, 2014, pp. 479-483.

- [10] J. Huang, Y.Q. Shi, Y. Shi, "Embedding image watermarks in DC components" (2000) IEEE Transactions on Circuits and Systems for Video Technology, 10 (6), pp. 974-979.
- [11] Z. Pan and H. Bolouri, "High speed face recognition based on discrete cosine transforms and neural networks", University of Hertfordshire, 1999.
- [12] E. Yen and L. Lin, "Rubik's cube watermark technology for grayscale images", Vol 37(6), pp 4033-4039, Jun. 2010.
- [13] B. L. Gunjal, and R. R. Manthalkar, "An overview of transform domain robust digital image watermarking algorithms", Journal of Emerging Trends in Computing and Information Sciences, vol. 2, no 1, CIS Journal 2010-2011.

A New Model for Pre-analysis of Network Traffic Using Similarity Measurement

Enas Ayman Al-Utrakchi

Zarqa University/Department of Computer Science, Zarqa, 13132, Jordan
e_utrakchi@yahoo.com

Mohammad Rasmi AL-Mousa

Zarqa University/Department of Software Engineering, Zarqa, 13132, Jordan
mmousa@zu.edu.jo

Abstract — Recently there has been a notable increases in the importance of intrusion detection systems' (IDS) ability to accurately identify new types of threats and attacks. However, limitations exist within IDS frequently rendering them inaccurate in detecting attacks; the primary reason for this has been false negative or false positive alarms. The purpose of this paper is to contribute to the enhanced detection of new attacks. Because of the developments mentioned above, improvements to IDS efficiency must be made to harden the system security. If current system security measures are left unresolved, increased frequency of system damage and/or crashes can be reasonably anticipated. Based upon experimental testing of sample attacks, this paper proposes a new method for updating a blacklist based on the extraction of patterns through comparison to known malicious activities and code, using similarity measurement against a predefined database and network traffic, after analysis and classification of anomalies. We demonstrate that intrusion black list modeling shows continued promise in making analysis of network traffic more efficient, and IDS critically more accurate.

Keywords—network security; intrusion; IDS; similarity measurement

I. INTRODUCTION

Through increases in use of Internet Technology (IT) and the ever increasing skill of attackers, previously unknown threats to network stability continue to abound. Because of this, the protection of Information Systems (IS) against malicious activities and attacks in networks has never been more critical. There already exist many software tools to strengthen network security, such as Anti-Virus, Firewalls, IDS, etc... This paper focuses on Intrusion Detection System (IDS); IDS is generally structured as second stage of defense. The design intent of IDS is to detect attacks and notify administrators whenever there is suspicious or abnormal traffic.

The most prominent limitation of IDS to date has been false readings of code and activity (either false positive or false negative) which result in inaccurate attack detection events. IDS is, therefore not perfectly equipped to detect new attacks. We suggest that if IDS designers strategically integrate techniques based in similarity theory IDS should realize increased detection accuracy.

Defensive security approaches such as Intrusion Detection System (IDS) and Intrusion Protection System (IPS) were developed to detect, prevent, and establish a perspective of network attacks. Intrusion must be analyzed more intensively to generate accurate data profiles

describing each attack and to establish a more suitable decision-making power within the IDS.

Intrusion analysis techniques (which were originally developed to enhance IDS) provide details about the traits of the attack and the behavior of the attacker. Analysis of attack intention is a prime example of the focus of intrusion analysis. Typical items of interest to intrusion analysis include IP addresses, ports, type of services and protocol, etc. The most common techniques for intrusion analysis depend on determining the features from the attack path as reported by [14-18]. The drawback of these techniques is that they are not suitable for large numbers of features. It has thus far been limited to handling specific types of features, and has yet to develop to an operationally feasible state, capable of efficiently presenting all the distinctives of an attack. With that limitation in mind, these techniques do work well with specific types of attacks, such as Distributed Denial of Service (DDOS) attacks.

In general, attack analysis is a critical and challenging task in security management [10]. The limited capability of security sensors and network monitoring tools make attack observation inaccurate and render the output incomprehensible.

This paper, presents a set of processes, shown in Fig. 1 that apply a similarity measurement to identify new attacks and add them efficiently to a blacklist. The proposed model

will be described in section 3 which defines the internal processes of the model. In section 4, we will explain the most salient ideas and issues in the current body of research. Finally, our experiments and discussion of findings are provided, based upon samples of statistical data describing known attack features.

II. RELATED WORKS

This section reviews literature on the network security, attacks and threats, intrusion detection systems and attack similarity theory.

Since the advent of the Internet, the number of LANs and personal computers has increased dramatically and this, in turn has given rise to a global, virtual environment, vulnerable to substantial security risks which arrive through networks. Firewall devices which impose an access control policy between two or more networks via software or hardware were a reaction to this new threat reality. Firewalls were primarily aimed at e-mail and Web surfing and sought to control inbound and outbound access to the Internet.

Network security is critically important in today's world because it secures all information passing through networked computers. Network security involves with all hardware and software functions, characteristics, features, operational procedures, accountability measures, access controls, administrative and management policy required to provide satisfactory level of safety for hardware, software, and information in a network [1].

Design and implementation of a network security model was presented in [2] using routers and firewall. Within that function, it also identified:

- 1) *the network security weaknesses in router and firewall network devices,*
- 2) *the type of threats arriving at entry points guarded by the firewall,*
- 3) *the system's responses to those threats, and*
- 4) *the method to prevent the attacks and hackers from accessing the network.*

The proposed model in [2] provides a checklist to use in evaluating whether a network is adhering to best practices in network security and data confidentiality. However, the model aims to protect the network from vulnerabilities, threats, attacks, configuration weaknesses and security policy weaknesses.

Generally, the authors of [3] started with the current situation of network security and analyzed its most influential elements to provide references the development of new models for securing computational property within networks.

An intrusion detection system detects various kinds of malicious network traffic and computer usage that cannot be detected by a conservative firewall. Thus, some researchers have focused on different types of attacks on IDS and given

descriptions of different attacks aimed at different protocols such as TCP, UDP, ARP and ICMP [4].

Whereas, the authors of [5] has concentrated on the design and develop the Intrusion Detection System for detecting Distributed Denial of Service (DDoS) Attacks in the network using Jpcap library in Java Programming language.

A new incremental hybrid intrusion detection system was proposed by [6]. This framework combines incremental misuse detection and incremental anomaly detection. The framework can learn new classes of intrusion that had not existed in previous data used for training incremental misuse detection. The framework has lower computational complexity so it is suitable for real-time or on-line learning.

Most intrusion analysis approaches are based on alert correlation techniques. These techniques are connected to network tools for assistance (such as IDS) to understand and analyze the intrusion event. The drawback of most of these techniques is that they are developed to prevent future attacks and minimize damage [11, 12, 13]. Thus, innovative methods and techniques are needed in the analysis of the attacks to increase the accuracy of the IDS through pre-analysis and the establishment of a robust intrusion blacklist in advance – this would provide the greatest help to IDS in their decision making.

Three novel algorithms based on the threshold algorithm were introduced in [7], it exploited the semantic properties of the new similarity measures to achieve the best performance in theory and practice.

Tanimoto based a similarity measure for host-based intrusions on a binary feature set for training and classification introduced by [8]. The k-nearest neighbor (KNN) classifier has been utilized to classify a given process either as normal or as an attack.

The Agile Similarity Attack Strategy (ASAS) model proposed by [9] heuristically identifies and monitors similar evidence between a new criminal case and others. The model uses a classification method based on a relation between attack evidence priorities with evidence group values, presented as a vector. Furthermore, the model uses a cosine similarity as a distance-based similarity measure (Metric Axioms) to improve the quality of decision making.

III. INTRUSION BLACKLIST MODEL

This section presents a new proposed model called the intrusion blacklist model. The proposed model adopts network capturing tools such as Wireshark, and Snort, as Network Intrusion Detection Systems (NIDSs). Network traffic is captured in the initial phase through network capturing tools, which normally produce a huge array of security data. This study will be empirical, using Snort, a free, open source system, whose core function is to monitor network traffic and detect attempts at intrusion. Where one or more tools within Snort will be used in each process (either in stand-alone status or collaboratively with tools

from other software packages which were tested working alongside Snort), there will be a series of processes which will lead to updating the intrusion blacklist. The proposed model includes five processes, i.e. collection, analysis, detection, classification, and similarity.

In the collection process, network traffic will be compiled using Snort which in turn converts network data into a coded form, pcap or log file, during real time network data traffic capture. The capture is then routed for analysis. Packets will be analytically read using tools include libraries such as libpcap to extract features according to signature and then stored to a database. This process is important because all subsequent processes will rely upon it; furthermore the useful selection of features at this stage will affect the work of snort.

During the detection process, operations are performed to detect abnormal activity and possible attacks in packets by matching features with rules ranked within a hierarchy – this process is one of the main stage in NIDS, since it takes responsibility assigning a proper action in response to each case intrusion: either log the packet or send an alarm notification. Naturally, if the packet is normal will be ignored. Once the full process is concluded, an output file will be created logging all alerts generated by NIDS (Snort). However, the rules of the NIDS can be controlled and amended by the user to improve result over time [19].

In the classification process, the NIDS generates many alerts some of which will be irrelevant. During this stage abnormal packets will be noted if categorized as a serious attack or not. Using k-nearest neighbor classifier in this way can detect a serious attack and minimize false alarms, then store them into database [20].

Finally, the similarity analysis process will be used to evaluate similar, abnormal packets through comparison with a predefined database, using Jaccard with Euclidean distance to estimate similarity of the most recently discovered attack features. This approach will improve IDS work through data refinement and the reduction of false alarms. The output will be an updated *Blacklist* database whose aim is to prevent future intrusive attacks [21].

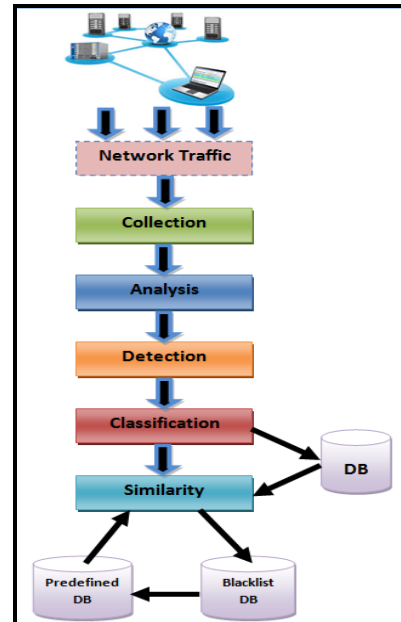


Fig. 1. Black list process model

IV. EXPERIMENTAL RESULTS AND ANALYSIS

This section evaluates the proposed model based on its analysis of a sample of 50 different types of attack. Each attack will be analyzed and extracted to identify 10 different type of features, such as, source IP address, destination IP address, port number, vulnerability, alert priority, time of life, type of protocol, type of service, ...etc.

Each feature will be numbered from 1-10, and it will be weighted for subsequent calculation based on criteria that rank the importance of each feature and the frequency of its occurrence.

Table I shows features' occurrence for each attack with a predefined attack database. The similarity percentage will be calculated using similarity measurement as a real-valued function that quantifies the similarity between the suspicious attack described in the Snort database and the observed attack.

TABLE I. EXAMPLE OF SIMILARITY OF THE ATTACK FEATURES

Feature #	Feature Similarity
1	12 %
2	30%
3	1%
4	0%
5	0.02%
6	0.22%

7	0%
8	0.32%
9	50%
10	2%

The evaluation of the results shows that the proposed intrusion blacklist model provides useful information and increases the possibility of detecting the real attack. Moreover, it helps IDSs eliminate the most similar features of the intrusions based on the similarity of attack features shown in Fig. 2; this helps improve the decision making process and the accuracy of the IDS.

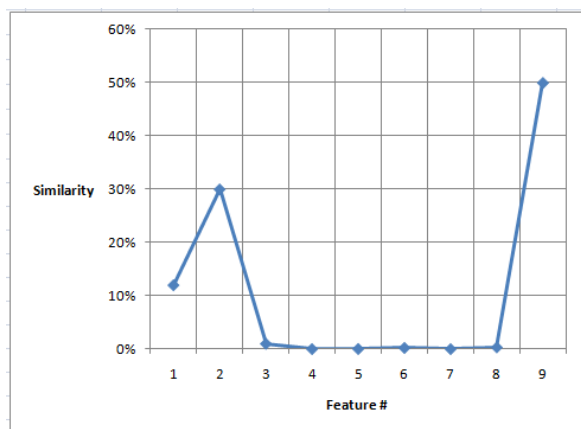


Fig. 2. Similarity of features of the attack

V. CONCLUSION AND FUTURE WORK

This paper proposes a new model to pre-analyze attacks during network traffic. The proposed model expects to make intrusion detection more accurate, an invaluable asset to IDS and all users. The proposed model will improve the quality of IDS decision making, in order to obtain clear information and achieve acceleration of the intrusion detection. Attack analysis is a critical and challenging task in network security management. Furthermore, features of attack recognition and analysis are an important research area in the field of network security. Obviously, to gain a higher ratio of intrusion detection, deeper analysis is desirable, as are more efforts to identify features of new attacks using suitable network security tools.

This paper is expected to conclude that most intrusion analysis approaches are based on alert correlation techniques which are used to understand and analyze the intrusion occurrence. Thus, the contribution of the research is anticipated to be the formulation of new, innovative methods and techniques aimed at increasing the accuracy of the IDS in order for it to be improved as a strong preliminary intrusion analysis tool capable of establishing a more reliable intrusion blacklist before actual attacks occur, and to thereby to help the IDS in its decision making.

REFERENCES

- [1] Chen S., Iyer R., and Whisnant K., "Evaluating the Security Threat of Firewall Data Corruption Caused by Instruction Transient Errors," In Proceedings of the 2002 International Conference on Dependable Systems & Network, Washington, D.C., 2002.
- [2] Alabady S., "Design and Implementation of a Network Security Model for Cooperative Network," International Arab Journal of e-Technology, Vol. 1, No. 2, June 2009
- [3] Zeng A., "Discussion and research of computer network security", Journal of Chemical and Pharmaceutical Research, 2014, 6(7):780-783
- [4] Anand A., Patel B., "an Overview on Intrusion Detection System and Types of Attacks It Can Detect Considering Different Protocols," International Journal of Advanced Research in Computer Science and Software Engineering, Vol.2, Issue 8, Aug. 2012
- [5] G. D. K., Rao CV, Singh M. k. and Kemal M., "Network-based IDS for Distributed Denial of Service Attacks," International Journal of Emerging Trends & Technology in Computer Science (IJETTCs), Vol. 3, Issue 1, Jan- Feb 2014
- [6] Rasoulifard A., Bafghi A. G., and Kahani M., "Incremental Hybrid Intrusion Detection Using Ensemble of Weak Classifiers, ser. Communications in Computer and Information Science. Springer Berlin Heidelberg, vol. 6, pp. 577–584. 2008
- [7] Hadjieleftheriou M., Chandel A., Koudas N., Srivastava D., "Fast indexes and algorithms for set similarity selection queries," in: Proceedings of the 24th International Conference on Data Engineering (ICDE '08), pp. 267–276.2008
- [8] Sharma A. and Lal S. P., "Tanimoto based similarity measure for intrusion detection system," Journal of Information Security, 2(4):195–201. 2011
- [9] Rasmi M. and Jantan A., "Asas: agile similarity attack strategy model based on evidence classification for network forensic attack analysis, AWERProcedia Information Technology & Computer Science 846-857. 2012
- [10] Qin, X. & Lee, W. , "Attack Plan Recognition and Prediction Using Causal Networks". Computer Security Applications Conference, 2004, 370-379.
- [11] Wei, W. & Thomas, E. D., "A Graph Based Approach Toward Network Forensics Analysis", ACM Trans. Inf. Syst. Secur., 2008, 12, 1-33.
- [12] Huang, M.-Y., Jasper, R. J. & Wicks, T. M., "A Large Scale Distributed Intrusion Detection Framework Based On Attack Strategy Analysis.", Computer Networks, 1999, 31, 2465-2475.
- [13] Damiano, B., Sandro, E. & Pieter, H. H., "Panacea: Automating Attack Classification for Anomaly-Based Network Intrusion Detection Systems.", Proceedings of The 12th International Symposium On Recent Advances In Intrusion Detection. Saint-Malo, France, Springer-Verlag, 2009, 1-20.
- [14] Wu, P., Zhigang, W. & Junhua, C., "Research On Attack Intention Recognition Based On Graphical Model.", Fifth International Conference On Information Assurance and Security, 2009. IAS '09.360-363.
- [15] Feng, J., Yuan, Z., Yao, S., Xia, C. & Wei, Q., "Generating Attack Scenarios for Attack Intention Recognition.", International Conference On Computational and Information Sciences. Chengdu, China, IEEE Computer Society, 2011, 272-275.
- [16] Hao, B., Kunsheng, W., Changzhen, H., Gang, Z. & Xiaochuan, J., "Boosting Performance In Attack Intention Recognition By Integrating Multiple Techniques.", Front. Comput. Sci China, 2011, 5, 109-118.
- [17] Peng, W., Yao, S. & Chen, J., "Recognizing Intrusive Intention and Assessing Threat Based On Attack Path Analysis." International Conference: Multimedia Information Networking and Security, 2009. MINES '09, 250-253.
- [18] Wang, Z. & Peng, W., "An Intrusive Intention Recognition Model Based On Network Security States Graph", 5th International

Conference On Wireless Communications, Networking and Mobile Computing, 2009. WICOM '09, 1-4.

- [19] Rehman R. U., “Intrusion Detection with SNORT: Advanced IDS Techniques Using SNORT, Apache, MySQL, PHP, and ACID”, Prentice Hall PTR , 2003.
- [20] Gabra H. N., Bahaa-Eldin A. M. and Korashy H., “ Classification of IDS Alerts with Data Mining Techniques”, International Journal of Electronic Commerce Studies Vol.5, No.1, pp.1-6, 2014.
- [21] Soldo F., Le A. and Markopoulou A.,” Predictive Blacklisting as an Implicit Recommendation System”, INFOCOM, Proceedings IEEE, PP 1 – 9,2010.

A New Authenticated Key Agreement Protocol

Kamal A. ElDahshan¹, Emad Masameer², AbdAllah A. Elhabshy³

Mathematics Department,
Faculty of Science, Al-Azhar University
Nasr City, Cairo, Egypt
¹dahshan@gmail.com
²emadmasameer@yahoo.com
³abdallah@azhar.edu.eg

Abstract—Authenticated key agreement protocols play a significant role in securing communications over public network channels (Internet). This paper proposes a new key agreement protocol based on factorization problem over nonabelian groups. Then it presents two different ways to provide mutual authentication for the proposed protocol; this paper presents a new authenticated key agreement protocol using fixed shared password and a new authenticated key agreement protocol using a digital signature. It also provides security analysis for the proposed two authenticated key agreement protocols.

Keywords— Cryptographic protocols; Key agreement protocol; Authentication; Digital signature; Security analysis

I. INTRODUCTION

To establish a secured communication, legitimate entities need to share a secret key. To limit the information available to the attacker, this key should be fresh each time they start a new communication (session). This can be done by using a key agreement protocol. A key agreement protocol is very important aspect in modern cryptography. Key agreement protocols [1] allow two or more entities in order to establish together a shared secret key. The value of this secret key is a function of the information contributed by the legitimate entities. In 1976, Diffie and Hellman proposed the first key agreement protocol [2] based on the public key cryptography. The security of the Diffie-Hellman protocol is based on the discrete logarithm problem (DLP) [3]. Nowadays, most secured communications use the Diffie-Hellman protocol in order to establish a secret key. If a polynomial algorithm is found to solve DLP, all these communications will be breakable at once. So, since Diffie-Hellman protocol, there are many attempts to construct a key agreement protocol based on other problems [4-12]. One of these problems is the factorization problem over non-abelian (non-commutative) groups [6]. Let $x \in G$, where G is non-abelian group, and $y = axb$, then the factorization problem is to find a, b satisfying $a^{-1}y = xb$. In this paper we propose a new protocol based on factorization problem over non-abelian groups. Then we present two directions in order to provide mutual authentication for our key agreement protocol. One of these directions use a fixed shared password between the legitimate entities, and the other use the digital signatures for authentication.

The rest of this paper is organized as follows. Section II proposes a new key agreement protocol. Section III presents

two authenticated key agreement protocols. One of them uses fixed shared password and the other uses a digital signature for authentication. Section IV provides the security analysis of our two authenticated key agreement protocols. Finally, Section V gives the conclusion and further work.

II. A NEW KEY AGREEMENT PROTOCOL

This section proposes a new key agreement protocol. This protocol is similar to Katvickis-Vitkus protocol [13] and Cho et al. protocol [6]. Let \mathbb{M} be the set of all $n \times n$ matrices over \mathbb{Z}_p , where p is a large prime. Let $\mathbb{M}_L, \mathbb{M}_R \subset \mathbb{M}$ such that $A_L B_L = B_L A_L \forall A_L, B_L \in \mathbb{M}_L$ and $A_R B_R = B_R A_R \forall A_R, B_R \in \mathbb{M}_R$. A singular matrix $P \in \mathbb{M}$ is publicly known. Also, $P A_L \neq A_L P \forall A_L \in \mathbb{M}_L$ and $P A_R \neq A_R P \forall A_R \in \mathbb{M}_R$. In this context our protocol can be described as follows:

1. Alice randomly selects two matrices $A_L \in \mathbb{M}_L$ and $A_R \in \mathbb{M}_R$. Then she computes and sends $Y_A = A_L P A_R \text{ mod } p$ to Bob.
2. Bob randomly selects two matrices $B_L \in \mathbb{M}_L, B_R \in \mathbb{M}_R$. Then he computes and sends $Y_B = B_L P B_R \text{ mod } p$ to Alice.
3. Alice computes the key $K_A = A_L Y_B A_R \text{ mod } p = A_L B_L P B_R A_R \text{ mod } p$.
4. Bob computes the key $K_B = B_L Y_A B_R \text{ mod } p = B_L A_L P A_R B_R \text{ mod } p$.

Since $A_L B_L = B_L A_L \forall A_L, B_L \in \mathbb{M}_L$ and $A_R B_R = B_R A_R \forall A_R, B_R \in \mathbb{M}_R$. Thus, Alice and Bob generate the same key $K = K_A = K_B = A_L B_L P B_R A_R \text{ mod } p$. Now, Alice and Bob can use the shared key in any cryptographic systems discussed in [14],[15] according to their needs.

According to [16], the complexity of matrix multiplication is $O(n^{2.37286})$, where n is the dimension of the square matrices. This mean the complexity of our protocol is $O(n^{2.37286})$.

Let $\mathbb{M}_C \subset \mathbb{M}$ be the set of all commutative matrices over \mathbb{Z}_p , i.e. $AB = BA \forall A, B \in \mathbb{M}_C$, and $\alpha, \beta \in \mathbb{N}$, where \mathbb{N} is the set of natural numbers. In our protocol, let $A_L = AP^{\alpha-1}$, $A_R = A^{-1}$, $A \in \mathbb{M}_C$, $B_L = BP^{\beta-1}$, and $B_R = B^{-1}$, $B \in \mathbb{M}_C$. Then we get an instance of Sakalauskas et al. schema [17]. Which also is similar to Eftekhari protocol [18] and Cho et al. protocol [6]. In this context, and instance protocol of Sakalauskas et al. schema can be described as follows:

1. Alice randomly selects an invertible matrix $A \in \mathbb{M}_C$ and $\alpha \in \mathbb{N}$. Then she computes and sends $Y_A = AP^{\alpha}A^{-1} \bmod p$ to Bob.
2. Bob randomly selects an invertible matrix $B \in \mathbb{M}_C$ and $\beta \in \mathbb{N}$. Then he computes and sends $Y_B = BP^{\beta}B^{-1} \bmod p$ to Alice.
3. Alice computes the key $K_A = A(Y_B)^{\alpha}A^{-1} \bmod p = A(BP^{\beta}B^{-1})^{\alpha}A^{-1} \bmod p = ABP^{\alpha\beta}B^{-1}A^{-1} \bmod p$.
4. Bob computes the key $K_B = B(Y_A)^{\beta}B^{-1} \bmod p = B(AP^{\alpha}A^{-1})^{\beta}B^{-1} \bmod p = BAP^{\alpha\beta}A^{-1}B^{-1} \bmod p$.

Since $AB = BA \forall A, B \in \mathbb{M}_C$, then Alice and Bob share the same key $K = K_A = K_B = ABP^{\alpha\beta}B^{-1}A^{-1} \bmod p$.

As mentioned, our protocol is similar to both Cho et al. and Eftekhari protocol. All of these protocols based on the factorization problem over the group of square matrices. They use the same security parameters (n, p) . Thus, our protocol is secure as Cho et al. and Eftekhari protocols, with the same values of security parameters. For more details see [6] and [18].

In our protocol if P is not a singular matrix and $AP = PA$ for some $A \in \mathbb{M}_L$ and $A \in \mathbb{M}_R$, then the attacker (Eve) can compute A_LA_R and deduce the shared key.

For more clarification, let $A_RP = PA_R \bmod p$, $A_RB_L = B_LA_R \bmod p$ and P is invertible matrix. Since P, Y_a and Y_b are publicly known. Then

1. $Y_a = A_LPA_R \bmod p = A_LA_RP \bmod p$ (since $A_RP = PA_R \bmod p$)
2. $A_LA_R = Y_aP^{-1} \bmod p$.
3. $K_E = A_LA_RY_a \bmod p = A_LA_RB_LPB_R \bmod p = A_LB_LA_RPB_R \bmod p$ (since $B_LA_R = A_RB_L \bmod p$)
4. $K_E = A_LB_LPA_RB_R \bmod p$ (since $A_RP = PA_R \bmod p$)
5. $K_E = A_LB_LPB_RA_R \bmod p$ (since $A_RB_R = B_RA_R \bmod p$)
6. $K_E = K \#$

III. AUTHENTICATED KEY AGREEMENT PROTOCOLS

Due the lack of authentication [19], the proposed protocol in Section II is vulnerable to the man-in-the-middle attack [3]. To provide an authentication, one can use a fixed shared password or a digital signature [20]. In this section we will use each of these ways in order to providing the authentication.

A. A New Authenticated Key Agreement Protocol Using Fixed Password (AKAP-Pwd)

Let Alice and Bob be two entities who share a secret password (Pwd) in advance. To use this password for authentication, they construct a secret matrix S from Pwd using a predetermined algorithm. Let H_K be a keyed hash function [21] and let \oplus_M be the bitwise XOR operation defined over matrices. The XOR operation of two matrices is done by XOR the corresponding coordinates of the entries of the two matrices. The matrices must be of the same order. In this context a new authenticated key agreement protocol can be described as follows:

1. Alice selects randomly two matrices $A_L \in \mathbb{M}_L$ and $A_R \in \mathbb{M}_R$ where $\mathbb{M}_L, \mathbb{M}_R \subset \mathbb{M}$. Then she computes and sends $Y_A = (A_LP A_R \bmod p \oplus_M S)$ to Bob.
2. Bob
 - a. selects randomly two matrices $B_L \in \mathbb{M}_L, B_R \in \mathbb{M}_R$.
 - b. computes the key $K_B = B_L(Y_A \oplus_M S)B_R \bmod p = B_LA_LP A_RB_R \bmod p$.
 - c. computes $Y_B = (B_LP B_R \bmod p \oplus_M S)$.
 - d. computes $V_B = H_{K_B}(Y_A || Y_B || \text{AliceID})$.
 - e. sends Y_B and V_B to Alice.
3. After receiving Y_B and V_B , Alice
 - a. computes the key $K_A = A_L(Y_B \oplus_M S)A_R \bmod p = A_LB_LP B_RA_R \bmod p$.
 - b. checks whether $V_B = H_{K_A}(Y_A || Y_B || \text{AliceID})$ or not. If it holds, Alice accepts the communication, otherwise Alice refuses the communication.
 - c. Alice computes and sends $V_A = H_{K_A}(Y_B || Y_A || \text{BobID})$ to Bob.
4. After receiving V_A , Bob checks whether $V_A = H_{K_B}(Y_B || Y_A || \text{BobID})$ or not. If it holds, Bob accepts the communication, otherwise he refuses it.

Since $A_LB_L = B_LA_L \forall A_L, B_L \in \mathbb{M}_L$ and $A_RB_R = B_RA_R \forall A_R, B_R \in \mathbb{M}_R$. Thus, Alice and Bob generate the same key $K = K_A = K_B = A_LB_LP B_RA_R \bmod p$.

B. A New Authenticated key Agreement Protocol Using Digital Signatures (AKAP-DS)

Let the entity's E public key be PuK_E and the entity's E private key be PdK_E . Let $\text{Sign}_E(M)$ be the E 's digital signature of the message M . And let $\text{Ver}_E(\text{Sign}_E(M)) = M$ be the verification algorithm of the digital signature $\text{Sign}_E(M)$. In this context, a new authenticated key agreement protocol can be described as follows:

1. Alice
 - a. selects randomly two matrices $A_L \in \mathbb{M}_L$ and $A_R \in \mathbb{M}_R$ where $\mathbb{M}_L, \mathbb{M}_R \subset \mathbb{M}$.
 - b. computes $Y_A = A_LP A_R \bmod p$.
 - c. computes the digital signature $S_A = \text{Sign}_{\text{Alice}}(H_t(Y_A))$, where t is a timestamp which has a unique value in each session.
 - d. Sends Y_A, t and S_A to Bob

2. Bob
 - a. ensures that t has never been used before, i.e. the value of t is new.
 - b. verifies S_A , i.e. checks whether $H_t(Y_A)? = \text{Ver}_{\text{Alice}}(S_A) = \text{Ver}_{\text{Alice}}(\text{Sign}_{\text{Alice}}(H_t(Y_A)))$ or not. If it holds, Bob proceeds with the protocol, otherwise he refuses the communication.
 - c. selects randomly two matrices $B_L \in \mathbb{M}_L, B_R \in \mathbb{M}_R$.
 - d. computes $Y_B = B_L P B_R \text{ mod } p$.
 - e. computes the key $K_B = B_L Y_A B_R \text{ mod } p = B_L A_L P A_R B_R \text{ mod } p$.
 - f. computes $S_B = \text{Sign}_{\text{Bob}}(H_{K_B}(Y_A || Y_B || \text{AliceID}))$
 - g. sends Y_B and S_B to Alice.
3. Alice
 - a. computes the key $K_A = A_L Y_B A_R \text{ mod } p = A_L B_L P B_R A_R \text{ mod } p$.
 - b. verifies S_B , i.e. checks whether $H_{K_A}(Y_A || Y_B || \text{AliceID})? = \text{Ver}_{\text{Bob}}(S_B)$ or not. If it holds, Alice accepts the communication, otherwise she refuses the communication.
 - c. computes and sends $H_{K_A}(Y_B || Y_A || \text{BobID})$ to Bob.
4. Bob checks whether $H_{K_A}(Y_B || Y_A || \text{BobID})? = H_{K_B}(Y_B || Y_A || \text{BobID})$ or not. If it holds, Bob accepts the communication, otherwise he refuses the communication.

Now, both Alice and Bob have the same secret key $K = K_A = K_B = A_L B_L P B_R A_R \text{ mod } p$.

In our AKAP-DS one can use any secure digital signature schema such as Yang-Liao schema [22] or Wu et al. schema [23] [24].

IV. SECURITY ANALYSIS

This section provides a security analysis of our two authenticated key agreement protocols mentioned in Section III. It is desirable for authenticated key agreement protocols to possess a numerous of security attributes [25, 26]. This section shows that our AKAP-Pwd and AKAP-DS possess these attributes. This section also shows that our AKAP-Pwd and AKAP-DS are immune to both passive and active attacks [27].

A. Security attributes

This section shows that our protocols possess the security attributes of Known-key security, Forward secrecy, Key compromise impersonation, Unknown key-share, Loss of information, Key control, and message independence [25, 26]. In what follows, Alice and Bob are two legitimate entities and Eve is an attacker.

Known-Key Security: In our two protocols both Alice and Bob choose new private matrices in each session. This means, Alice and Bob construct a new key in each session. So, knowing an old session key does not affect the security of the current key.

Forward secrecy: In our two protocols if the long-term keys (password in our AKAP-Pwd and private keys in our AKAP-DS) are revealed by an attacker, there is no effect of the previous session keys. This is because in our protocols the long-terms are used only for authentication purposes, and do not affect the value of the key.

Key-compromise impersonation: As any protocol uses a secret shared password for authentication, in our AKAP-Pwd if Eve knows the secret password she can impersonate Alice to Bob and impersonate Bob to Alice. In other words, our AKAP-Pwd does not provide the attribute of “key-compromise impersonation”. While in our AKAP-DS, if Eve covers the private key of a legitimate entity (say Alice), then Eve can impersonate Alice but Eve cannot impersonate others to Alice. In other words, our AKAP-DS guarantees the attribute of “key-compromise impersonation”.

Unknown key-share: Both AKAP-Pwd and AKAP-DS are designed in such ways that make it impossible for Eve to fool a legitimate entity (say Bob) to share a session key with Alice without his knowledge. In our AKAP-Pwd, the password is shared only between Alice and Bob, which means Eve neither can establish a session key with Bob as Alice nor with Alice as Bob. In our AKAP-DS, each entity has her/his own private key to prove her/his identity, i.e. Eve cannot impersonate a legitimate entity.

Loss of information: Loss of information that is not usually available to Eve does not affect the security of our protocols in other sessions. For example if Eve knows A_L or/and A_R in some session(s), she cannot know the secret key that has been (will be) established in any other session(s). This is because in each session both Alice and Bob choose new random matrices in order to construct a new key.

Key control: In both AKAP-Pwd and AKAP-DS, neither Alice nor Bob can control the value of the key. This is because the value of the key is a function of the information supplied by each of Alice and Bob. So, both AKAP-Pwd and AKAP-DS guarantee the attribute of key control.

Message independence: The flows in AKAP-Pwd and AKAP-DS are deliberately independent. The attribute of message-independence is important, it prevents many possible attacks such as “on-line/off-line password guessing attack”, and replay attacks.

B. Passive and Active attacks

There are two main types of attacks, passive attacks and active attacks. In passive attacks, an attacker (Eve) can only eavesdrop the communication between Alice and Bob. Meanwhile, Eve analyzes the transformed message in order to compute the secret key or any other useful information (guessing the password or cryptanalyzing the key agreement protocol). Our protocol is based on the factorization problem which is a generalization of the conjugacy search problem [28]. To break the protocol using the brute force attack, Eve needs to check out all possible keys. As discussed in [29] for a similar protocol, the protocol is secured if $p = 251$ and $n = 32$ which make the key

length equals to 8192 bits, i.e. the key space equal to 2^{8192} . If Eve has a computer with CPU speed $1\text{THz} = 10^{12}\text{Hz}$ which does not exist until now, she needs time $\cong 2^{8127}$ year $\cong 2^{8097}$ billion year (since since $10^{12} * 60 * 60 * 24 * 365.25 \cong 2^{65}$) to check every possible key. This is much larger than the age of the universe (the age of the universe is $\cong 13.8$ billion year [30]). In other words, our protocols are immune to brute force attack.

In active attacks, an attacker (Eve) can capture, modify, and resend messages or even initiate and construct new messages. There are many types of active attacks such as modification attacks, replay attacks and off-line password guessing attacks. In what follows, we will discuss the security of our protocols under each type of these active attacks.

Modification attacks: In modification attacks, Eve captures and modifies the messages (flows) in order to modify the shared key. Consider the scenario in which Eve tries to modify Y_A to Y'_A . Then, in our AKAP-Pwd (Section III.A), Alice will not accept the communication as soon as she checks V_B in step 3. Also, in our AKAP-DS (Section III.B) Bob will not accept the communication as soon as she checks the Alice's digital-signature in step 2. Now, consider the scenario in which Eve tries to modify Y_B to Y'_B . Then, Alice will refuse the communication as soon as she checks V_B (step 3) in our AKAP-Pwd, or as soon as she checks Bob's digital-signature (step 3) in our AKAP-DS.

Replay attacks: Each of our protocols is deliberately designed in a way to ensure that it is impossible for an attacker (Eve) to replay any message without the knowledge of the legitimate entities.

Off-line password guessing attack: Off-line password attack could be done by a passive or an active attacker. In off-line password attack, the attacker (Eve) tries to find the shared password between the legitimate entities and prove the correctness of this password. Since the flows (messages) are independent in our AKAP-Pwd, there is no way to find the secret shared password using the transmitting messages.

V. CONCLUSION AND FURTHER WORK

This paper proposed a new authenticated key agreement protocol. Then it presented two authentication methods. The first uses a fixed shared password (AKAP-Pwd) and the second uses a digital signature (AKAP-DS). Then this paper provided security analysis for both AKAP-Pwd and AKAP-DS. It proved that our authenticated protocols guarantee the desirable security attributes for authenticated key agreement protocols. Moreover, the paper showed that both AKAP-Pwd and AKAP-DS are immune to passive and active attacks.

This work will be enhanced by presenting a new reference schema for authenticated key agreement protocols [31].

REFERENCE

[1] A. J. Menezes, S. A. Vanstone, and P. C. V. Oorschot, Handbook of Applied Cryptography: CRC Press, 1997.

- [2] W. Diffie and M. E. Hellman, "New directions in cryptography," IEEE Transactions on Information Theory, vol. 22, pp. 644-654, 1976.
- [3] W. Stallings, Cryptography and Network Security: Principles and Practice, 6th ed.: Prentice Hall Press, 2014.
- [4] V. M. Sidelnikov, M. A. Cherepnev, and V. Y. Yashchenko, "Systems of open distribution of keys on the basis of noncommutative semigroups," Acad. Sci. Dokl. Math, vol. 48, pp. 384-386, 1993.
- [5] K. Ko, S. Lee, J. Cheon, J. Han, J.-s. Kang, and C. Park, "New Public-Key Cryptosystem Using Braid Groups," in Advances in Cryptology — CRYPTO 2000. vol. 1880, M. Bellare, Ed., ed: Springer Berlin Heidelberg, 2000, pp. 166-183.
- [6] S. Cho, K.-C. Ha, Y.-O. Kim, and D. Moon, "Key Exchange Protocol Using Matrix Algebras and Its Analysis," Journal of Korean Mathematical Society, vol. 42, pp. 1287-1309, 2005.
- [7] V. Shpilrain and A. Ushakov, "The Conjugacy Search Problem in Public Key Cryptography: Unnecessary and Insufficient," Applicable Algebra in Engineering, Communication and Computing, vol. 17, pp. 285-289, 2006/08/01 2006.
- [8] E. Sakalauskas, A. Katvickis, and G. Dosinas, "Key Agreement Protocol over the Ring of Multivariate Polynomials," Information Technology and Control, vol. 39, pp. 51-54, 2010.
- [9] H. K. Pathak and M. Sanghi, "Public key cryptosystem and a key exchange protocol using tools of non-abelian group," International Journal on Computer Science and Engineering, vol. 2, pp. 1029-1033, 2010.
- [10] V. Ottaviani, A. Zaroni, and M. Regoli, "Conjugation as public key agreement protocol in mobile cryptography," in Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on, 2010, pp. 1-6.
- [11] J.-J. Climent, P. Navarro, and L. Tortosa, "Key exchange protocols over noncommutative rings. The case of," Int. J. Comput. Math., vol. 89, pp. 1753-1763, 2012.
- [12] D. Kahrobaei, C. Koupparis, and V. Shpilrain, "Public Key Exchange Using Matrices Over Group Rings," Groups Complexity Cryptology, vol. 5, pp. 97-115, 2013.
- [13] A. Katvickis and P. Vitkus, "Key Agreement Protocol Using Elliptic Curve Matrix Power Function," in Advanced Studies in Software and Knowledge Engineering, ed: Institute of Information Theories and Applications FOI ITHEA, 2008, pp. 103-106.
- [14] D. S. A. Elminaam, H. M. A. Kader, and M. M. Hadhoud, "Evaluating The Performance of Symmetric Encryption Algorithms," International Journal of Network Security, vol. 10, pp. 213-219, 2010.
- [15] D. S. A. Minaam, H. M. Abdual-Kader, and M. M. Hadhoud, "Evaluating the Effects of Symmetric Cryptography Algorithms on Power Consumption for Different Data Types," International Journal of Network Security, vol. 11, pp. 78-87, 2010.
- [16] F. L. Gall, "Powers of tensors and fast matrix multiplication," presented at the Proceedings of the 39th International Symposium on Symbolic and Algebraic Computation, Kobe, Japan, 2014.
- [17] E. Sakalauskas, P. Tvarijonas, and A. Raulynaitis, "Key Agreement Protocol (KAP) Using Conjugacy and Discrete Logarithm Problems in Group Representation Level," Informatica (lithuanian Academy of Sciences), vol. 18, pp. 115-124, 2007.
- [18] M. Eftekhari, "A Diffie-Hellman Key Exchange Using Matrices Over Non Commutative Rings," Groups, Complexity and Cryptology, vol. 4, pp. 167-176, 2012.
- [19] J. E. Canavan, Fundamentals of Network Security: Artech House, 2001.
- [20] B. A. Forouzan, Introduction to Cryptography and Network Security: McGraw-Hill Higher Education, 2008.

- [21] M. Bellare, R. Canetti, and H. Krawczyk, "Keying Hash Functions for Message Authentication," in *Advances in Cryptology — CRYPTO '96*, vol. 1109, N. Kobitz, Ed., ed: Springer Berlin Heidelberg, 1996, pp. 1-15.
- [22] F.-Y. Yang and C.-M. Liao, "A Provably Secure and Efficient Strong Designated Verifier Signature Scheme," *International Journal of Network Security*, vol. 10, pp. 220-224, 2010.
- [23] W. Wu, Y. Mu, W. Susilo, and X. Huang, "Server-Aided Verification Signatures: Definitions and New Constructions," in *Provable Security*, vol. 5324, J. Baek, F. Bao, K. Chen, and X. Lai, Eds., ed: Springer Berlin Heidelberg, 2008, pp. 141-155.
- [24] Z. Wang, L. Wang, Y. Yang, and Z. Hu, "Comment on Wu et al.'s Server-aided Verification Signature Schemes," *International Journal of Network Security*, vol. 10, pp. 238-240, 2010.
- [25] S. Blake-Wilson, D. Johnson, and A. Menezes, "Key Agreement Protocols and Their Security Analysis," presented at the Proceedings of the 6th IMA International Conference on Cryptography and Coding, 1997.
- [26] S. Blake-Wilson and A. Menezes, "Authenticated Diffie-Hellman Key Agreement Protocols," presented at the Proceedings of the Selected Areas in Cryptography, 1999.
- [27] R. A. MOLLIN, *An Introduction to Cryptography*, Second ed.: Taylor & Francis, 2007.
- [28] D. Moldovyan and N. Moldovyan, "A New Hard Problem over Non-commutative Finite Groups for Cryptographic Protocols," in *Computer Network Security*, vol. 6258, I. Kottenko and V. Skormin, Eds., ed: Springer Berlin Heidelberg, 2010, pp. 183-194.
- [29] E. Sakalauskas, N. Listopadskis, and P. Tvarijonas, "Key Agreement Protocol (KAP) Based on Matrix Power Function," in *Sixth International Conference on Information Research and Applications Varna, Bulgaria, 2008*, pp. 92 - 96.
- [30] C. L. Bennett, D. Larson, J. L. Weiland, N. Jarosik, G. Hinshaw, N. Odegard, K. M. Smith, R. S. Hill, B. Gold, M. Halpern, E. Komatsu, M. R. Nolte, L. Page, D. N. Spergel, E. Wollack, J. Dunkley, A. Kogut, M. Limon, S. S. Meyer, G. S. Tucker, and E. L. Wright, "Nine-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Final Maps and Results," in *Cosmology and Nongalactic Astrophysics*, ed. NY, United States: Cornell University, 2013, pp. 1-177.
- [31] A. A. Elhabshy., "A Generalized Form for Key Agreement Protocols.," PhD thesis, Faculty of Science - Department of Mathematics Al-Azhar University, Cairo - Egypt, expected summer 2015, unpublished.

The Generalised Secured Mobile Payment System Based on ECIES and ECDSA

Ehab M. Alkhateeb

Faculty of Science and Information Technology: Al Zaytoonah University of
Jordan
Amman, Jordan
ehabalkh@gmail.com

Mohammad A. Alia

Faculty of Science and Information Technology: Al Zaytoonah University of
Jordan
Amman, Jordan
dr.m.alia@zuj.edu.jo

Adnan A. Hnaif

Faculty of Science and Information Technology: Al Zaytoonah University of
Jordan
Amman, Jordan
dr.adnan_hnaif@zuj.edu.jo

Abstract— Mobile payment system is defined as an electronic payment method, also it is defined as mobile money transfer and mobile wallet. Since mobile payment has been generated to be an attractive alternative for the traditional payments systems such as credit cards. In this paper Elliptic curve cryptography is used to mobile payment system. Meanwhile, the proposed mobile payment system includes three main processes: Authentication process, Member recognition process, and Payment process. Moreover, Elliptic Curve Integrated Encryption Scheme ECIES and Elliptic Curve Digital Signature Algorithm ECDSA cryptographic protocols have been applied to enhance the security of the proposed mobile payment system. However, the proposed system is secure, easy and straightforward payment process. As well as, USSD technology is used in this system for PIN authentication process with high security performance.

Keywords— ECC, ECIES, ECDSA, Mobile Payment System, and Cryptography.

I. INTRODUCTION

Mobile payment is defined as payment for products or services between two parties for which a mobile device, such as a mobile phone, plays a key role in the realization of the payment [1]. Nowadays mobile phones are spreading widely through social communities, and becoming a replacement for laptops and desktop PCs. The user demands for convenient and intelligent ways in which to make payments for goods and services using a mobile phone is creating exciting opportunities for those organizations that are part of the mobile payment ecosystem [2]. However, Mobile payments facing critical security issues with the rise of identity fraud, and illegal access to confidential data, such as credit card details. Furthermore, the cryptanalysis and attacking, protocols speed, and performance evaluation are the core elements in building a secure mobile payment system. Therefore, this paper focuses its attention on these concerns by presenting a mobile payment which is based on public key cryptography. The assessment of security for the proposed mobile system is based on the

strength of proposed cryptographic algorithm, the selected key size, the performance and the speed of the proposed system. This approach is rather a replacement or a merge for classical way to pay using credit card and the current mobile payment methods.

II. RELATED WORKS

A. Cryptography

Cryptography is a cornerstone of the modern electronic security technologies used today to protect valuable information resources on intranets, extranets, and the Internet. Cryptography is the science of providing security for information [3]. Cryptography have many algorithms that can be categorized into two main types based on the nature of key, namely secret and public keys. The secret key or non-public key cryptosystem only need one key(secret key) to encrypt and decrypt the data between the sender and the recipient, while public-key cryptosystems comes in more difficult approach, it

consists of two keys, the public key which is used to encrypt the data and private key for decryption. Public key based on key exchange protocol rises above the difficulties faces by the secret key cryptosystem. This is because key management is much easier with the help of a key exchange protocol such as Diffe-Hellman [4].

B. RSA (Rivest, Shamir, Adelman) protocol

The RSA protocol [5] is one of most widely used public key cryptography algorithms. The algorithm was invented in 1977 by Ron Rivest, Adi Shamir, and Leonard Adleman. The RSA algorithm relays on large integers and prime testing, its mathematical basis is the Euler theorem, the security of RSA depends on the difficulty of factoring larger integer [6]. However, RSA have many disadvantages, and with time passing its being replaced with more efficient algorithms, the following are some disadvantages of RSA cryptosystem [6]:

- Fake public key.
- Complexity of key creation.
- Security need to be proofed.
- Slow of the speed.

C. RSA Digital Signature (RSA DS)

In the RSA algorithm for encryption and decryption process uses public key to encrypt and private key to decrypt as mentioned previously. The RSA Digital Signature (RSA DS) uses the private key to generate a signature and the public key is used to verify that signature [6].

D. ECC Cryptography

Elliptic Curve Cryptography (ECC) nowadays having a lot of attention, due to its small key size for encryption, decryption and digital signature, beside entered a wide use in 2004 and 2005. ECC was discovered by Nael Kobitz [7], and Victor S. Miller [8]. The ECC schemes are public-key mechanisms that provide the same functionality as RSA schemes. However, their security is based on the hardness of a different problem, namely the Elliptic Curve Discrete Logarithmic Problem (ECDLP) [9]. The adoption for ECC makes it a competitive to RSA, since it can reach the same security level with smaller key size, smaller key size means less computation time and high performance speed. ECC can be used in various areas, as encryption algorithm like ECIES which is a replaceable for RSA cryptosystem, or key exchange protocol such as ECDH, or as a digital signature such as ECDSA which recently being used intensively through the internet to provide integrity and non repudation for messages. Elliptic curve protocols depend on ECDLP, which it is assumed that finding the discrete logarithm of a random elliptic curve element with respect to a publicly known base point is infeasible.

E. ECIES cryptosystem

One of the most efficient encryption and decryption based on elliptic curve is ECIES. ECIES is a public-key cryptosystem

[10]. ECIES proposed by Abdalla, Bellare, and Rogway in [11] and [12]. As its name properly indicates, ECIES is an integrated encryption scheme which uses the following functions [10]:

- Key Agreement (KA): Function used for the generation of a shared secret by two parties.
- Key Derivation Function (KDF): Mechanism that produces a set of keys from keying material and some optional parameters.
- Encryption (ENC): Symmetric encryption algorithm.
- Message Authentication Code (MAC): Data used in order to authenticate messages.
- Hash (HASH): Digest function, used within the KDF and the MAC functions

To apply encryption and decryption using ECIES between Alice (sender) and Bob (receiver) must do the following [3]:

- a. The previous Cryptographic functions.
- b. Elliptic Curve domain parameters (p, a, b, G, n, h) for a curve over prime field or $(m, f(x), a, b, G, n, h)$ for a curve over binary field.
- c. Keys generation:
 - Bob's public key : V (Bob generates it as follows: $V = v.G$, where v is the private key he chooses at random: $v \in [1, n-1]$)
 - Alice's public key: U (Alice generates it as follows: $U = u.G$, where u is the secret key she chooses at random: $u \in [1, n-1]$)
 - Optional shared information (parameters): $S1$ and $S2$.
- d. Encryption: To encrypt a message m , Alice does the following:
 1. Derives a shared secret: $S = Px$, where $P = (Px, Py) = u.V$ (and $P \neq 0$)
 2. Uses KDF to derive a symmetric encryption and a MAC keys: $K_{ENC} \parallel K_{MAC} = KDF(S \parallel S1)$;
 3. Encrypts the message: $c = ENC(K_{ENC}; m)$;
 4. Computes the tag of encrypted message and $S2$: $d = MAC(K_{MAC}; c \parallel S2)$;
 5. Outputs $U \parallel c \parallel d$.
- e. Decryption: to decrypt the ciphertext, Bob does the following:
 1. Cryptogram $U \parallel c \parallel d$.
 2. Derives the shared secret: $S = Px$, where $P = (Px, Py) = v.U$ (it is the same as the one Alice derived because $P = v.U = U.v.G = u.V.G = u.V$, or outputs failed if $P = 0$;
 3. Derives keys the same way as Alice did: $K_{ENC} \parallel K_{MAC} = KDF(S \parallel S1)$;
 4. Uses MAC to check the tag and outputs failed if $d \neq MAC(K_{MAC}; c \parallel S2)$;
 5. Uses symmetric encryption scheme to decrypt the message $m = ENC^{-1}(K_{ENC}; c)$

Figure 1 shows ECIES encryption and decryption process between Alice, and Bob.

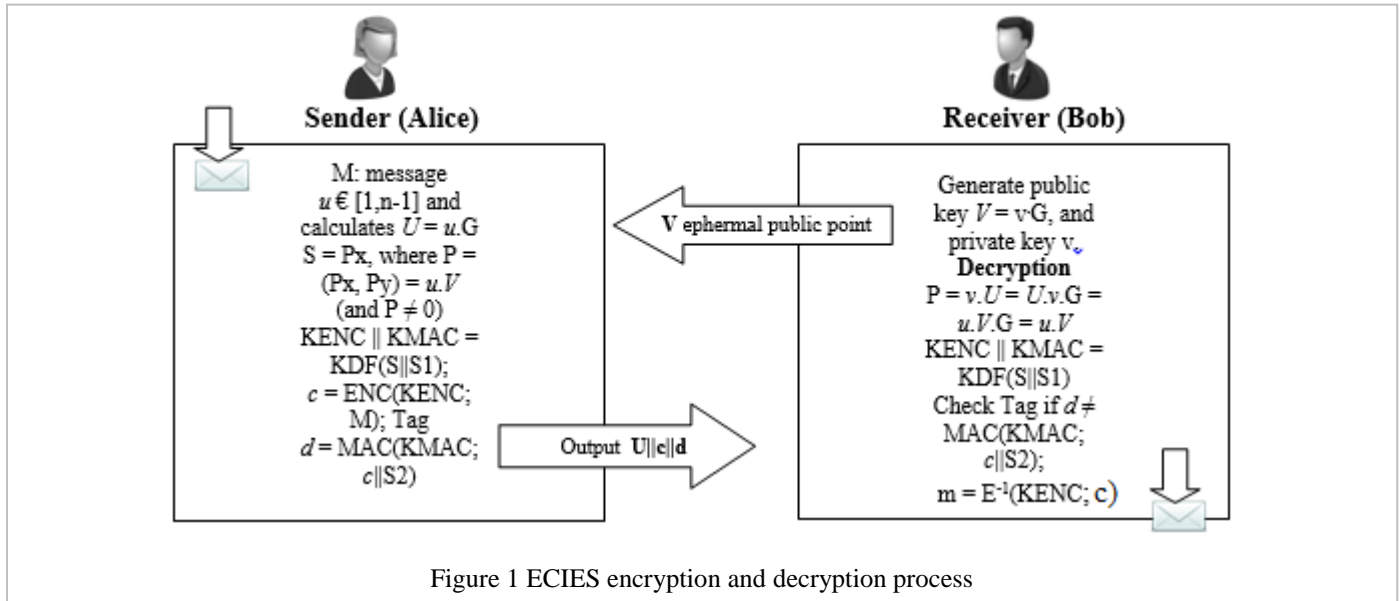


Figure 1 ECIES encryption and decryption process

F. Elliptic Curve Digital Signature Algorithm (ECDSA)

The ECDSA is a variant of the Digital Signature Algorithm (DSA) that operates on elliptic curve groups, for sending a signed message from A to B, both have to agree up on Elliptic Curve domain parameters. The domain parameters are defined in section Elliptic Curve Domain parameters. Sender 'A' have a key pair consisting of a private key d_A (a randomly selected integer less than n , where n is the order of the curve, an elliptic curve domain parameter) and a public key $Q_A = d_A * G$ (G is the generator point, an elliptic curve domain parameter) [3].

G. Identity-Based Cryptography (IBC)

The identity based concept depends on user's identifier information, such as phone number, e-mail, IP address etc., which replaces the digital certificates and to use public key for encryption or signature verification. Thus, this reduces the cost dramatically for establishing and managing the public key authentication framework known as Public Key Infrastructure (PKI) [13]. However, IBC facing major challenges as the following:

- Key ESCROW problem.
- Authentication is not sufficient.
- No support for certificates.

H. End to end encryption (E2EE)

The E2EE is a concept used to secure data on flight from one device to another, in a pure networking it means to secure data between two endpoints unlike the client-server architecture. Figure 2 shows one of the variants E2EE namely POS to Acquirer Encryption (P2AE) where data are being encrypted before it being sent to the bank acquirer [14].

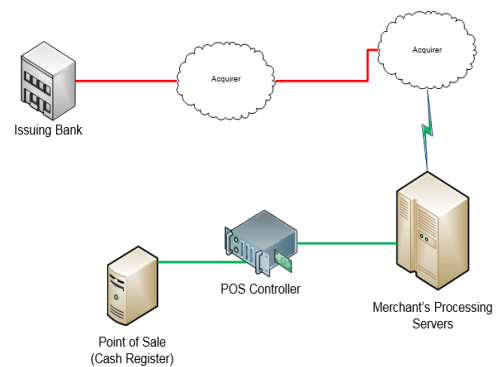


Figure 2 P2AE model [14].

I. USSD technology

The USSD technology is commonly used in banking, mobile polling, security systems and education, USSD is standard GSM technology supported by all GSM handsets. It is session-based and supports longer message content. Secure and cost-effective, sessions can be initiated by both end-users and enterprises [15]. USSD messages are more secured than SMS, data are protected when transmitted with default GSM security, and can be categorized into two types:

1. Push messages: Message received to the user that acquires a response.
2. Pull messages: Message initiated by the user by dialing then to acquire response.

USSD can be used for security improvements for banks or any other service that requires authentication, such as PIN entrance or one time password for a transaction authentication [16].

III. MOBILE PAYMENT SYSTEMS

There have been a number of deployments of mobile payments worldwide across the spectrum of proximity and remote payment. The following are some of these deployments:

A. A secure mobile payment service (SEMOPS)

Among the popular mobile payments systems, SEMOPS is a mobile payment solution that is capable of supporting micro, mini payments, and a universal solution, being able to function in any channel, including mobile, Internet and Point of Sale (POS), SEMOPS depends on securing transaction data by using RSA cryptosystem [17]. However, SEMOPS uses data center to control the flow of data, which requires many cooperate processes with different mobile operators and banks to accept the transaction. Thus, making it a much higher cost.

B. Google Wallet

Google Wallet is a popular mobile payment system developed by Google that allows users to store different payment cards in their mobile phones, Google Wallet uses NFC-technology to initiate payments on any pay-pass terminal at checkout [18]. NFC-technology is provided by verifone, and also provide E2EE protecting customer confidential data on transmittion.

C. Heartland mobile payment

Heartland appraich the same as in google wallet, both requires an NFC, and a wallet at POS, also the payment process steps almost the same. However, Heartland depend on Voltage security which provide an identity-base encryption, which also adopted by others mobile payment systems for E2EE to protect cardholder and sensitive authentication data throughout the payment acquiring (e.g. bank) network [19].

D. Paybox

Paybox a mobile payment system founded in 1999, Paybox spread in many Europe countries. Paybox Germany bank-

centric mobile payment system connects with different banks to authorize mobile payments through customer bank accounts [20]. Paybox McDonalds in France have the same approach as in Google wallet, which depend on VeriFone, as part of its security is VeriShield, VeriShield provides robust security by protecting mobile payment transaction details on POS at rest or transmit through E2EE between merchant and payment processor or bank which uses RSA PKI relies on digital certificates as well as public and private cryptographic keys to secure information exchange, the unencrypted data will never exposed to thieves [21].

E. ECC for securing payment system

This mobile payment system is based on ECC cryptography, the Diffie-Hellmann key exchange protocol DHKEP is implemented by sending the public key of each party over insecure channel, also digital signature is used for authentication. The Cryptographic application used covers one of the mobile payment security requirements framework elements, namely: the application service provider which guarantees a secure end-to-end Communication and non-repudiation [22]. However, Elliptic Curve Diffie Hellman ECDH protocol lacks authentication. Thus, digital signature algorithm is used to overcome this issue.

F. Mobile payment method based on RSA

This system is based on RSA cryptosystem, also adopt SMS technology for PIN entrance [23]. However, RSA as mentioned previsoulsy facing many problems. Furethermore, SMS technology also facing problems related to security such as Identity fraud and man in the middle attacks.

G. Challenges facing current mobile payments

Mobile payments systems facing many challenges as mentioned previously, the main challenges can be classified as the following:

- Security :
 - a. Cryptographic services lacks efficiency, and authentications needed.
 - b. Vulnerability of many attacks, such as identity fraud or man in the middle attacks.
- Low supported phones: Low support of applied technologies such as NFC.
- Extra hardware or software for the customer.

IV. THE PROPOSED SYSTEM

The proposed secured mobile payment system (SMPS) mainly divided into three process, these are:

1. Authentication Process.
2. Client Recognition Process.
3. Payment Process.

The following are stakeholders of the proposed system:

- Customer.

- Merchant : act as mediator between the customer and the bank
- The Bank: Central point in the mobile payment, acting on behalf of payment processor, and manages the payment operation.
- Mobile Operator: act as service provider.

A. Authentication (Getting Service) Prcoess

The first process in the proposed system is the authentication process, where the bank, mobile operator takes the main role in this process, before the client can be able to use this service, the client must first register. As illustrated in Figure 3 step 1 the client meet with the bank employee and request the service, the bank employee checks if the client did provide the mobile phone number that he or she will use at the time of payment, if the number was not provided through their account, the bank employee will ask for it. As shown in Figure 3 step 2 the bank should acknowledge the mobile operator to authorize this service through sending the customer information. The mobile operator will respond with a notification message to the client of a successful registration after approving this service as show in Figure 3 step 3. The bank will complete the registration by giving the client the PIN as shown in Figure 3 step 4. Following these four steps the bank will generate a computed ECIES public-key $Bpu1$, and private-key $Bpv1$, also a computed ECDSA public-key

$Bpu2$, and private-key $Bpr2$ (refer to Figure 3 step 5) and then pass the two public keys $Bpk1$, and $Bpk2$ to the market server (refer to Figure 3 step 6). Following that the market will generate a computed ECIES public-key $Mpu1$, and private-key $Mpr1$, also a computed ECDSA public-key $Mpu2$, and private- $Mpr2$ (refer to Figure 3 step 7) and then pass the two public keys $Mpu1$, $Mpu2$ to the bank (refer to Figure 3 step 8). At this stage the customer can shop in the market and pay through the mobile payment service.

B. Client recognition process

When the client enters the market for shopping (figure 4 step 1), the market server should test the customer either registered to the service or not (figure 4 step 2 and 3). If so a notification message will be sent to acknowledge that he or she can pay through the mobile payment service (figure 4 step 4). Otherwise, the customer will be ignored (figure 4 step 5).

C. Payment process

The final process in the proposed system is the payment process. The payment process divided into two phases, these are the payment phase and the payment confirmation phase.

- Payment phase:

The payment process starts when the client would like to buy

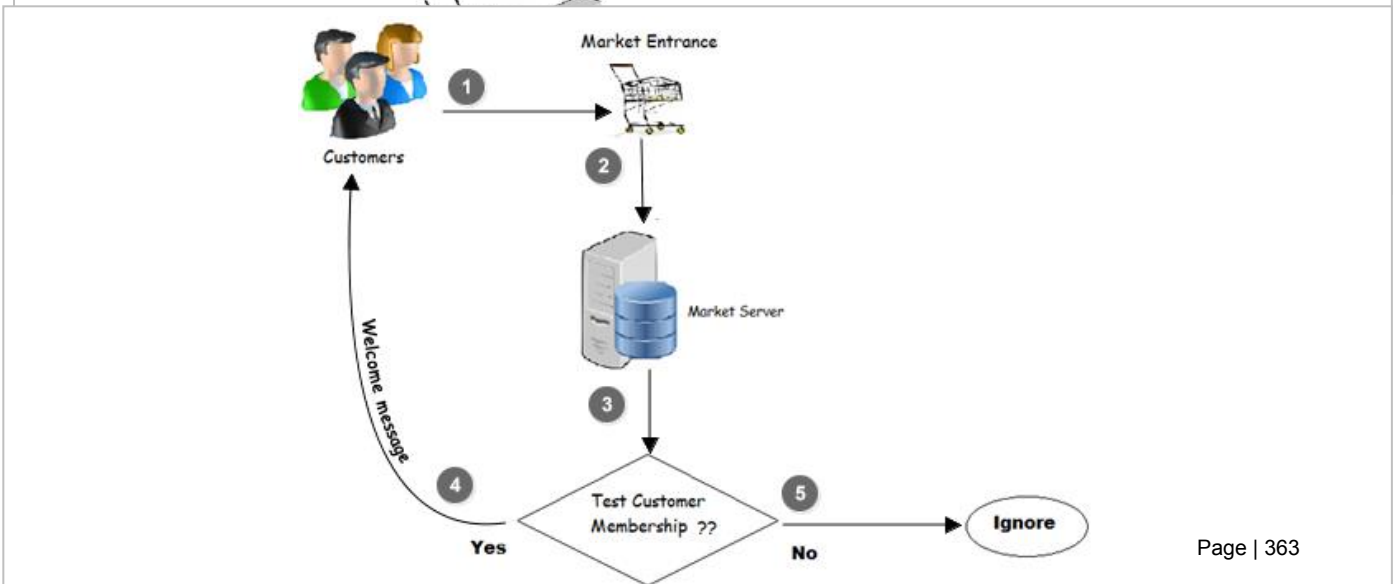
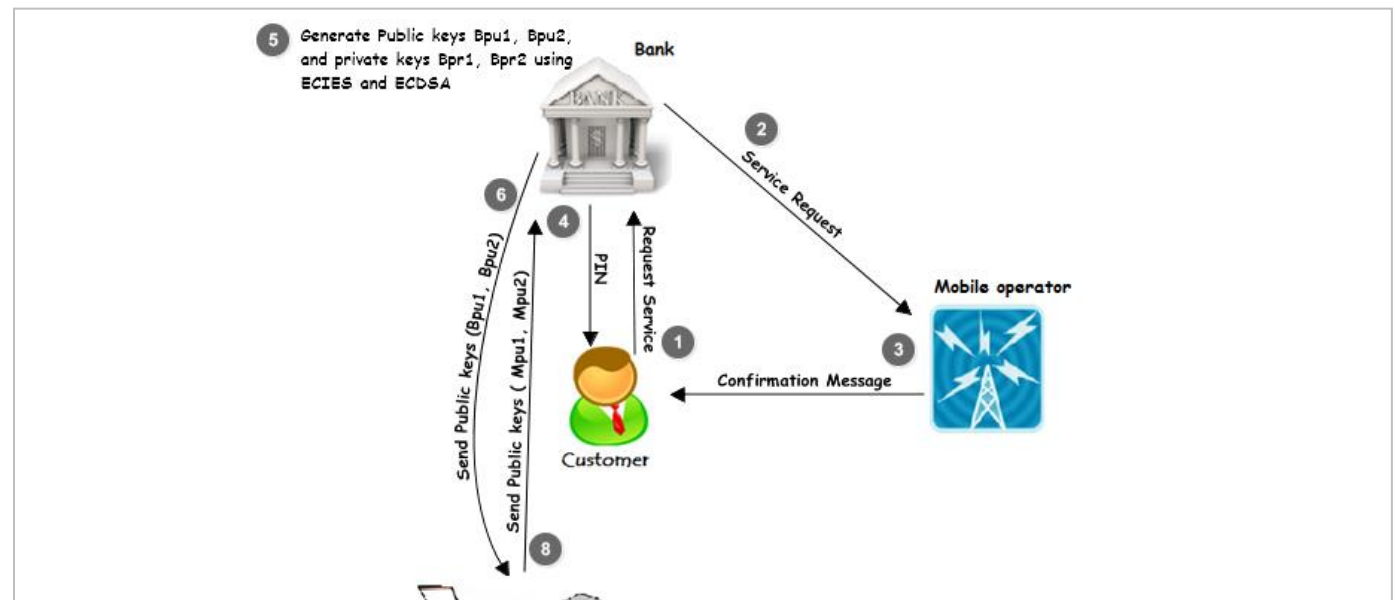


Figure 4 Client recognition process.

via mobile phone. First, the client have the choice to pay through mobile phone, Credit Card, or in cash as shown in figure 5 step1. At this stage if the client chooses the mobile payment, the client must pass the mobile phone number to the merchant to start processing the payment (figure 5). After that the following steps will be implemented securely, the customer membership will be checked through market server (Figure 5 step 2). If the client was a member, then the following steps will take place. The market server should contact the bank to provide the client information and the market information, the client information includes the amount to pay, and mobile phone number, while the market information includes the market ID and password, meanwhile the bank should make sure that the amount is available as shown in figure 5 step3.

sent to the bank. When the bank receives the ciphertext (encrypted text) two cases will be implemented. First, the ECIES is implemented to decrypt the data using bank private key ($Bpr1$), secondly the ECDSA is implemented to verify the signature using merchant public key ($Mpu2$) as shown in figure 5 step 4. After a successful implementation for the previous two cases the bank tests for valid amount and also checks for valid merchant ID and password as illustrated in figure 5 step 5, then upon successful checking process the bank should send the result to the mobile operator. After that, if the amount is valid as shown in figure 5 step 6 the mobile operator sends a USSD message to the client for payment confirmation, the client should respond to the message with the PIN as shown in figure 5 step 7 and 8. In figure 5 step 9 shows the deduction of the amount from the mobile operator to the bank.

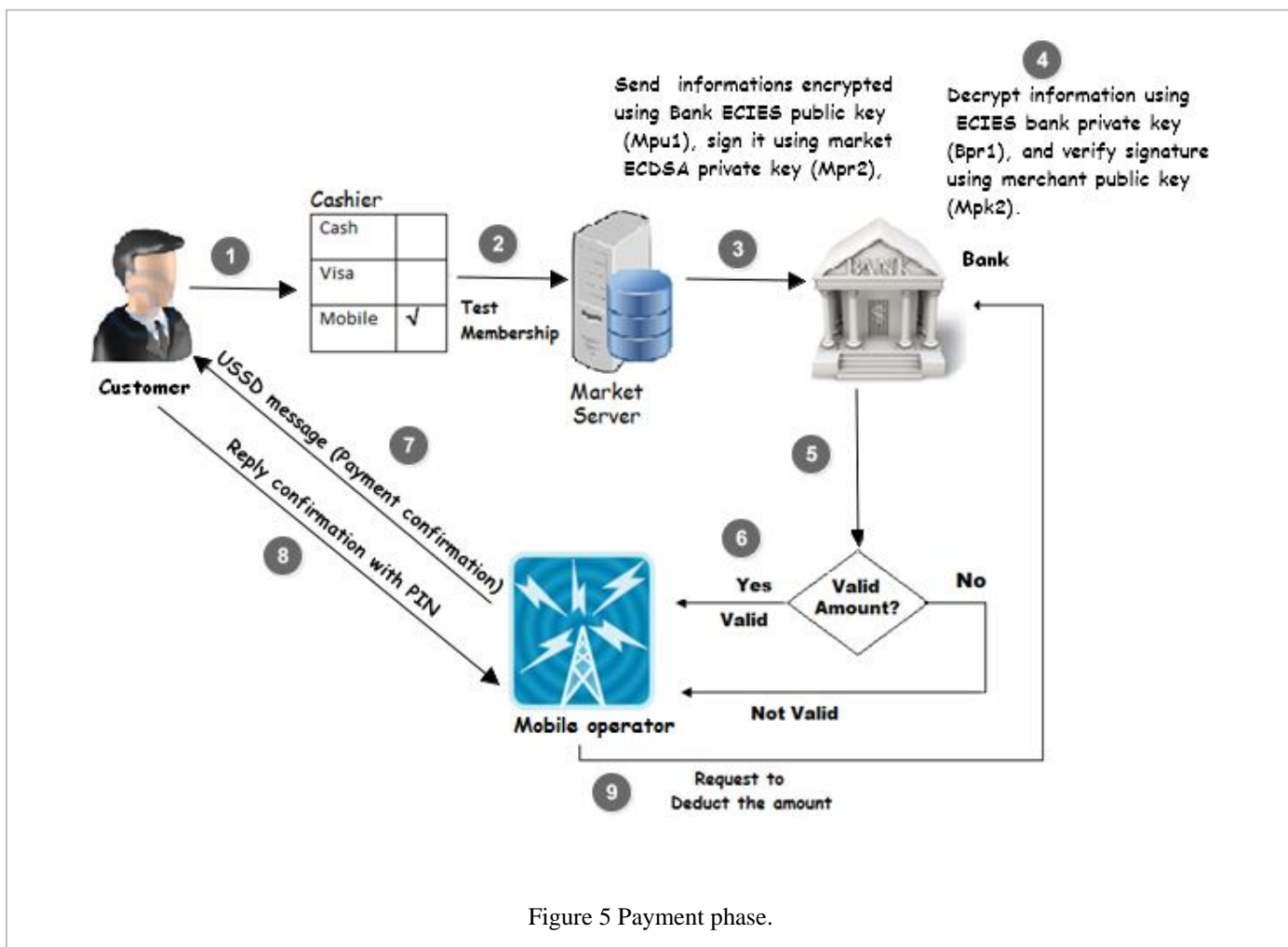


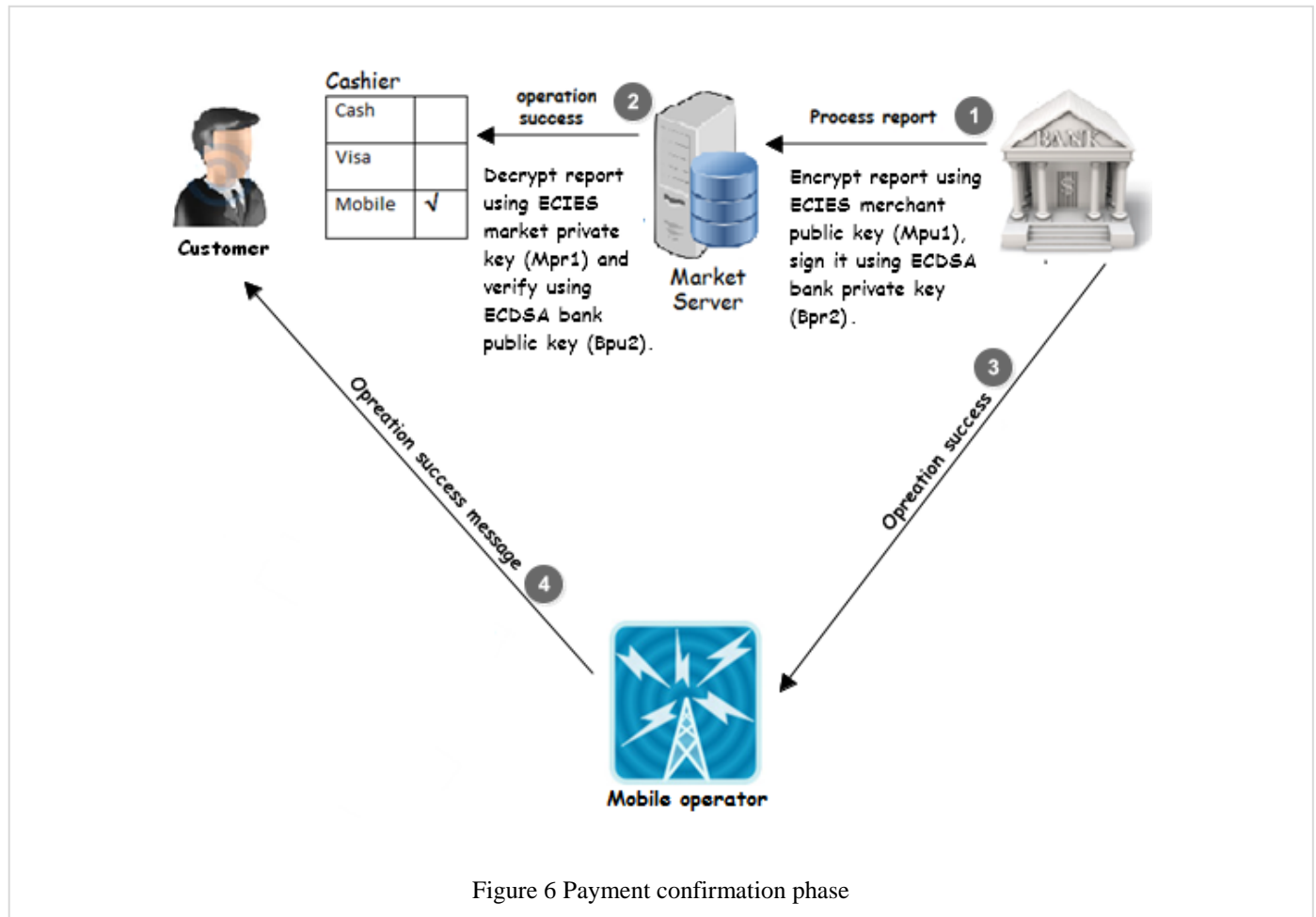
Figure 5 Payment phase.

Therefore, the ECIES, and ECDSA algorithms are implemented to encrypt and authenticate the customer information alongside the market information, this can be done using the bank public keys ($Bpu1$) to ciphertext (encrypted text) and using the merchant private key ($Mpr2$) to generate a signature. After that all these information will be

• Payment Confirmation phase:
 Following the payment phase it comes the payment confirmation phase in the payment process as shown in figure 6. This phase start through sending a report confirming that the payment process is executed successfully from the bank to

the merchant server (figure 6 step1). This report is encrypted by implementing the ECIES algorithm using the merchant server public-key ($Mpu1$), and signed using the bank private-key ($Bpr2$).

($Mpr1$) and verify signature using the bank public-key ($Bpu2$) as shown in figure 6 step 2. At this stage, the bank sends an operation success notification to the mobile operator (figure 6 step 3). Lastly, the mobile operator sends an operation success



message to the client (figure 6 step 4).

D. The advantages of the proposed system :

The following are main advantages of the proposed system:

- Secured under theft, the PIN is secured using GSM channel through USSD push message. USSD does not store messages in the phone like SMS which rises the risk of fraud. Furthermore, it's secured from Man in the Middle attacks.
- Provide high security level with high performance using ECIES and ECDSA applied on end-to-end encryption between merchant and the bank.
- No extra hardware or software needed for the customer in order to use the service.
- Wide range support of mobile phones, classical and smart ones.
- Easy and straightforward payment experience for customers.
- A replacement for the traditional credit card payments.

The market server receives the encrypted report and decrypt it by applying the ECIES algorithm using the market private-key

TABLE I. MBILE PAYMENT CHARACTERISTICS

Payment systems	Characteristics		
	Cryptographic algorithms	Possibility of Identity Theft	Extra Hardware requirements
Google Wallet	RSA and RSA DS	High	Yes
SEMOPS	RSA and RSA DS	High	Yes
Heartland	Boneh Franklin (IBC)	High	Yes
paybox	RSA	High	Yes
ECC for securing mobile payment system	DHKEP	-	-
Payment method based on RSA	RSA encryption	High	No
Proposed system (SMPS)	ECIES and ECDSA	Secured	No

V. EXPERIMENTAL RESULTS

In this section we discuss our experimental results of SMPS by applying RSA, ECIES, and ECDSA cryptosystems. The code implemented in windows 7 environment with Intel core due processor, using flexiprovider, bouncycastle APIs and NetBeans IDE. Flexiprovider is a powerful toolkit for the Java Cryptography Architecture (JCA/JCE) [24]. The implementation includes keys generation process and encryption, decryption processes and sign, verify processes. Figure 7 shows the computation time for RSA and RSA DS keys generation of length 1024-bit, 2048-bit, and 3072-bit with ECIES and ECDSA keys of length 160-bit, 224-bit and 256-bit corresponding to symmetric security level of 80-bit, 112-bit, and 128-bit, the simulation shows a faster keys generation for ECIES and ECDSA. ECIES and ECDSA public keys is a point on the curve, and the private keys are generated randomly, this gives them an advantage for generating both keys in a very short time. Figure 8 shows the computation time for encryption and decryption processes for RSA-1024, RSA-2048, and RSA-3072 keys and figure 9 shows encryption and decryption processes for ECIES-160, ECIES-224, and ECIES-256 keys, the simulation shows relatively faster encryption and decryption for ECIES, because ECIES relying on the hardness of the discrete logarithm problem in elliptic curve groups. It consumes low computation time for encryption and decryption and uses small key size.

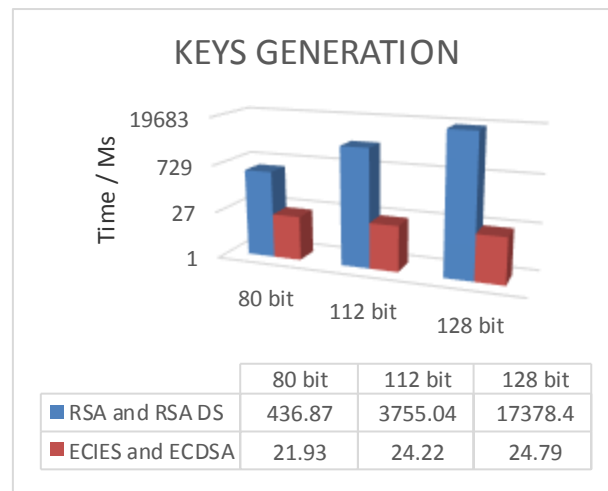


Figure 7 RSA and ECIES Keys generation

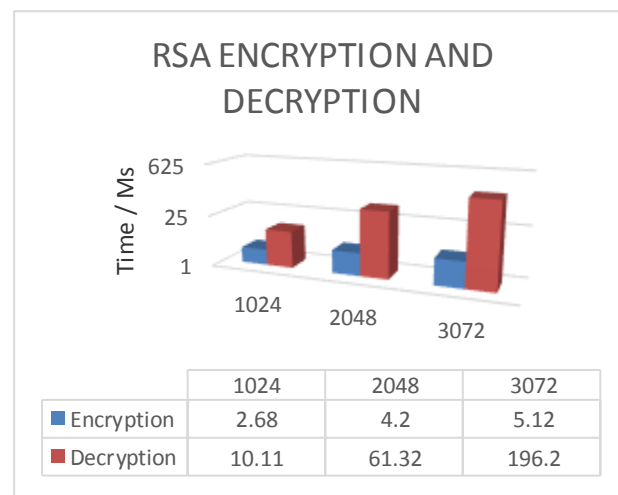


Figure 8 RSA encryption and decryption process.

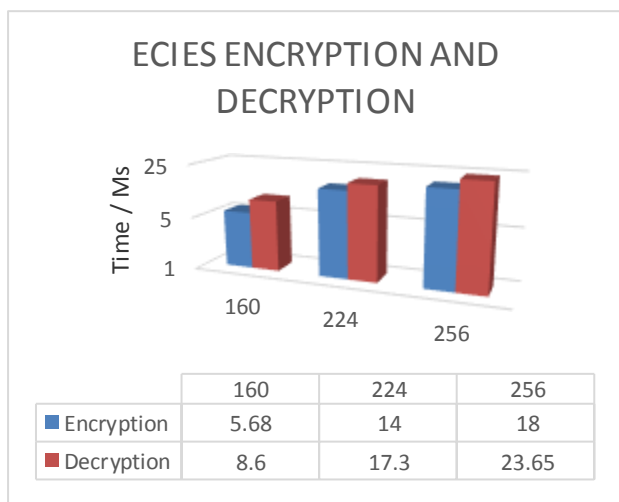


Figure 9 ECIES encryption and decryption process.

Figure 10 shows the average time for signing and verifying processes between RSA DS and ECDSA using the previous keys length. And finally figure 11 and 12 shows the performance evaluation for recommended security keys life time of keys lengths for ECIES and ECDSA of 224-bit, and 256-bit with RSA and RSA DS of 2048-bit, and 3072-bit.

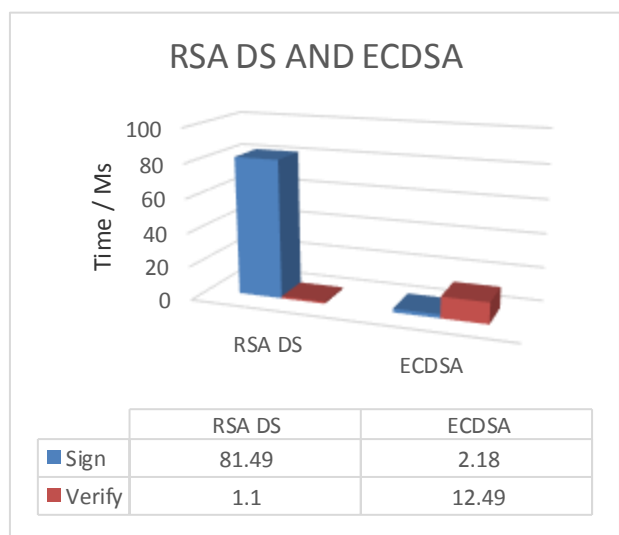


Figure 10 Average time for signing and verifying between RSA DS and ECDSA.

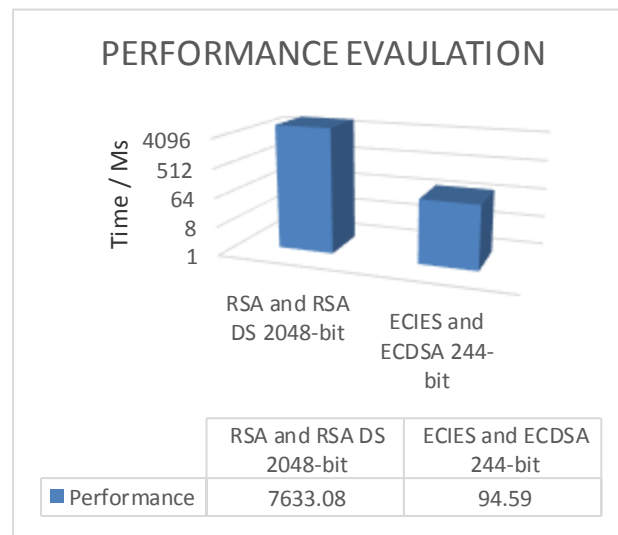


Figure 11 Performance evaluation 1.

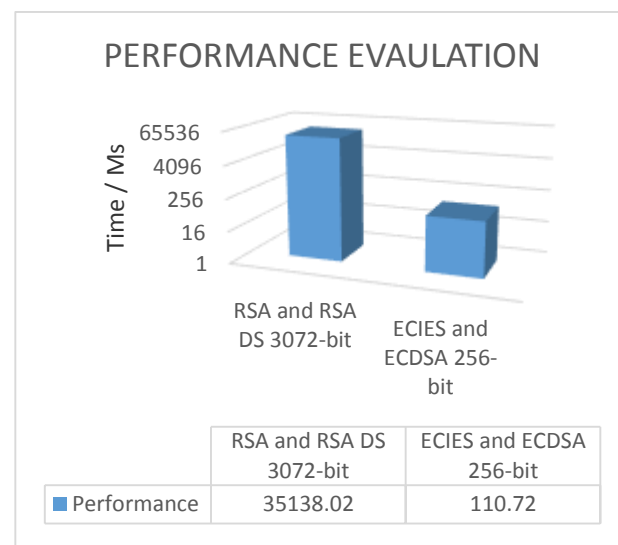


Figure 12 Performance evaluation 2.

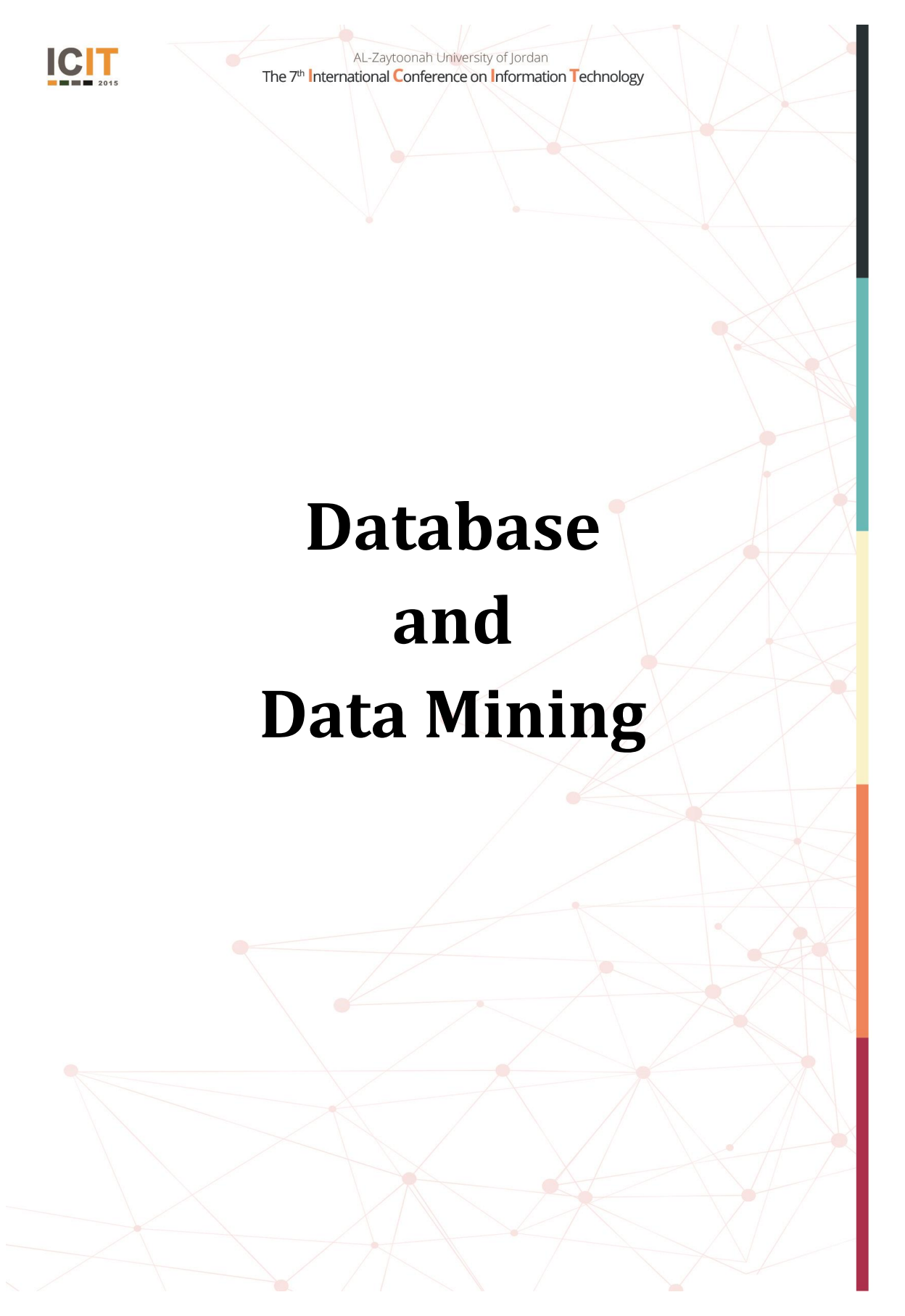
VI. CONCLUSION

This paper shows the possibility of establishing mobile payment system based on Elliptic Curve Cryptography. The security issue of the system depends on the hardness of elliptic curve discrete logarithm problem. Elliptic curve cryptography is actually adopted to its efficiency and requires small key size comparing to RSA cryptosystem with the same security level. This system is considered as an alternative method to displace the current traditional payment systems, in order to insure the security and confidentiality issues. As well as, the propose system needs minimum requirements such as; mobile phone, mobile operator, and market server.

REFERENCES

- [1] L. Bailly and B. Van der Lande, "Breakthroughs in the european mobile payment market," Atos Oringin, 2007.
- [2] H. Wilox, "Checkout the mobile payment opportunity," Juiper research, 2007.
- [3] R. Markan and K. Gurvinder, "Literature Survey on Elliptic Curve Encryption Techniques," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, no. 9, pp. 906-909, 2013.
- [4] M. A. Alia and A. B. Samsudin, "New key exchange protocol based on Mandelbrot and julia fractal sets," International Journal of Computer Science and Network Security, vol. 3, no. 9, pp. 906-909, 2007.
- [5] R. A. Rivest, A. Shamir and L. Adleman, "A Method for Obtaining Digital Signatures and Public-Key Cryptosystems," Communications of the ACM, vol. 21, no. 2, pp. 120-126, 1978.
- [6] NaQi, Wei Wei, J. Zhang, W. Wei , J. Zhao, J. Li, P. Shen, X. Yin, X. Xiao and J. Hu , "Analysis and Research of the RSA algorithm," Asian network for scientific Information, vol. 12, no. 9, pp. 1818-1824, 2013.
- [7] N. Koblitz, "Elliptic Curve Cryptosystems," Mathematics of Computation, vol. 48, no. 177, p. 203-209, 1987.
- [8] V. Miller , "Use of elliptic curves in Cryptography," Springer-Verlag, vol. CRYPTO '85, no. LNCS 218, pp. 417-426, 1986.
- [9] D. S. Kumar, C. Suneetha and A. ChandrasekhAR , "Encryption of data using elliptic curve over finite fields," International Journal of Distributed and Parallel systems, vol. 3, no. 1, pp. 301-308, 2012.
- [10] V. G. Martinez, L. H. Encinas and C. San, "A survey of the elliptic curve inegrated encryption scheme," Journal of Computing Science and Engineering, vol. 2, no. 2, pp. 7-13, 2010.
- [11] M. Abdalla, M. Bellare and P. Rogaway, "DHAES: An encryption scheme based on the Diffie-Hellman problem," submission to IEEE <http://grouper.ieee.org/groups/1363/P1363a/contributions/dhaes.pdf> , 1998.
- [12] M. Abdalla, M. Bellare and P. Rogaway, "DHIES: An encryption scheme based on the Diffie Hellman problem," unpublished. <http://www.cs.ucdavis.edu/~rogaway/papers/dhies.pdf> , 2001.
- [13] Girish and Phaneendra , "Identity-Based Cryptography and Comparison with traditional Public key Encryption: A Survey," International Journal of Computer Science and Information Technologies, vol. 5, no. 4, pp. 5521-5525, 2014.
- [14] B. Williams, "Will End to End Encryption Save Us All," Brandonwillams Secure Business Growth, 2010.
- [15] Infobip, "USSD Interactive Services," 2015. [Online]. Available: <http://www.infobip.com/services/ussd/>. [Accessed 10 2 2015].
- [16] Tekutiev, "USSD Interactive Services," 2013. [Online]. Available: <http://eyeline.mobi/blog/author/nikita-tekutiev/>. [Accessed 12 2 2015].
- [17] S. Karnouskos, A. Vilmos and A. Ram, "SeMoPS: A Global Secure Mobile Payment Service," 2005.
- [18] Mashable, 2014. [Online]. Available: <http://mashable.com/category/google-wallet/> . [Accessed 12 2 2015].
- [19] V. Security, "Infinite Peripherals Partners with Voltage Security to Enhance Mobile Payment Data Protection," 2015. [Online]. Available: <http://www.voltage.com/company/news/press-room/pr140210-infinite-peripherals-partners-with-voltage-security-to-enhance-mobile-payment-data-protection/>. [Accessed 25 1 2015].
- [20] O. Santolalla, "Mobile payment as key factor for mobile commerce success," Helsinki University of Technology, 2008.
- [21] VeriFone, "Verishield Total Protect," 2014.
- [22] A. Tohari, "Elliptic Curve Cryptography for Securing Payment Sysmtem," in ICTS, Bali, 2013.
- [23] A. Hnaif and M. Alia, "Mobile payment method based on public-key cryptography," International Journal of computer networks & communications, vol. 7, no. 2, pp. 81-92, 2015.
- [24] FlexiProvider. [Online]. Available: <https://www.flexiprovider.de/>. [Accessed 23 10 2014].

Database and Data Mining



A Formal Mathematical Semantics of Advanced Operations of Multiset Table Algebra

Iryna Glushko

Applied Mathematics, Informatics and Educational Measurement Department
Nizhyn Gogol State University
Nizhyn, Ukrain
iryna.glushko@ndu.edu.ua

Abstract— Multiset table algebra is considered. The notion of a table specified using the notion of a multiset (or bag). A signature of multiset table algebra is filled up with new operations such as inner and outer joins, semi-join and aggregate operations. A formal mathematical semantics of these operations is defined. The special element NULL is inserted in the universal domain for a define of the outer join.

Keywords—relation databases; multiset table algebra; inner joins; outer joins; semi-join, aggregate operations

I. INTRODUCTION

There are many applications the most peculiar feature of which is multiplicity and repeatability data. For example, these are sociological polls of different population groups, calculations on DNA and others. Commercial relational database systems are almost invariably based on multisets instead of sets. In other words, tables are in general allowed to include duplicate tuples. For example, the data model of SQL is relational in nature, as well as the relevant operations. However, unlike relational algebra, the tables manipulated by SQL are not relations, but, rather, multisets. The reason for this peculiarity is twofold. First, this is due to a practical reason: since SQL tables may be very large, duplicate elimination might become a bottleneck for the computation of the query result. Second, SQL extends the set of query operators by means of aggregate functions, whose operands are in general required to be multisets of values.

The relational model is based on the sets of tuples, i.e. it does not allow duplicate tuples in a relation [1]. So, naturally there is a need to expand possibilities of relational databases due to use of multisets. This problem was also considered in [2-5]. However this question requires specification and extension because in the specified works the due attention isn't paid to operations of inner and outer joins, semijoin and aggregate operations of multiset table algebra.

II. MULTISSET: BASIC DEFINITIONS

Let's introduce the basic concepts of multisets in terms of monograph [5].

Definition 1. A multiset α with basis U is a function $\alpha: U \rightarrow N^+$, where U is an arbitrary set, $N^+ = \{1, 2, \dots\}$ is the set of natural numbers without zero.

Let D be a universe of element of multiset bases, and then power set $P(D)$ – a universe of multiset bases. Let α be a multiset with basis $U_\alpha = \text{dom}\alpha$. Here $\text{dom}\alpha$ is the range of definition of multiset as a function.

Definition 2. A characteristic function of multiset α is a function $\chi_\alpha: D \rightarrow N$, the values of which are specified by the following piecewise schema: $\chi_\alpha(d) = \begin{cases} \alpha(d) & \text{if } d \in \text{dom}\alpha, \\ 0, & \text{else;} \end{cases}$ for all $d \in D$.

Definition 3. An empty multiset \emptyset_m is a multiset a characteristic function of which is a constant function, value of which is everywhere equal zero.

Definition 4. A rank of finite multiset α is a sum of duplicate elements of its basis $\|\alpha\| = \sum_{d \in \text{dom}\alpha} \alpha(d)$; wherein $\|\emptyset_m\| = 0$.

Let's introduce a binary relation inclusion over multisets.

Definition 5. Multiset β is included in multiset α ($\beta \preceq \alpha$), if $U_\beta \subseteq U_\alpha$ & $\forall d(d \in U_\beta \Rightarrow \beta(d) \leq \alpha(d))$. Directly from definition follows that this relation is a partial order.

The 1-multisets are the multisets whose range of values is an empty set or a single-element set $\{1\}$. These multisets are the analogues of ordinary sets.

The operations over multisets are defined in terms of characteristic functions in monograph [5]. There are operations of multiset union \cup_{All} , intersection \cap_{All} , difference \setminus_{All} , which build multisets of general view. The Cartesian product of multiset \otimes , the operation $Dist(\alpha)$, which build 1-multiset, and analog of a full image for multisets are defined too.

III. MULTISET TABLE ALGEBRA: BASIC DEFINITIONS

Among the two sets that are considered, A is the set of attributes and D is the universal domain.

Definition 6. An arbitrary (finite) set of attributes $R \subseteq A$ is called the scheme.

Definition 7. A tuple of the scheme R is a nominal set on pair R, D . The projection of this nominal set for the first component is equal to R .

The set of all tuples on scheme R is designated as $S(R)$ and the set of all tuples is designated as S .

Definition 8. A table is a pair $\langle \psi, R \rangle$, where the first component ψ is an arbitrary multiset basis of which $\Theta(\psi)$ is an arbitrary set (in particular infinite) of tuples of the scheme R and other component R is a scheme of table.

Thus, a certain scheme is ascribed to every table. The set of all table on scheme R is designated as $\Psi(R)$ and the set of all table is designated as $\Psi = \bigcup_R \Psi(R)$.

The notation $Occ(s, \psi)$ denotes the number of duplicate tuple s in the multiset ψ . Let's agree a multiset to write down as $\{s_1^{n_1}, \dots, s_k^{n_k}\}$, where $n_i = Occ(s_i, \psi)$, $i = 1, \dots, k$, and $\Theta(\psi) = \{s_1, \dots, s_k\}$ is a basis of the multiset ψ .

Definition 9. The multiset table algebra is the algebra $\langle \Psi, \Omega_{P, \Xi} \rangle$, where Ψ is the set of all tables, $\Omega_{P, \Xi} = \{ \cup_{All}^R, \cap_{All}^R, \setminus_{All}^R, \sigma_{p, R}, \pi_{X, R}, \otimes_{R_1, R_2}, Rt_{\xi, R}, \sim_R \}^{p \in P, \xi \in \Xi}_{X, R, R_1, R_2 \subseteq A}$ is the signature, P, Ξ are the sets of parameters.

The operations of signature $\Omega_{P, \Xi}$ are defined in [6].

IV. THE ADVANCED OPERATIONS

The advanced operations include inner and outer joins, semijoin, aggregate operations.

A. Inner Joins

There are four kinds of inner join operations Cartesian join, natural join, join using attributes A_1, \dots, A_n and join on predicate p . Let's define them.

Definition 10. The Cartesian Join of table on scheme R_1 and table on scheme R_2 , moreover $R_1 \cap R_2 = \emptyset$, is a binary parametric operation Cj_{R_1, R_2} of the form

$$Cj_{R_1, R_2} : \Psi(R_1) \times \Psi(R_2) \rightarrow \Psi(R_1 \cup R_2),$$

$$\langle \psi_1, R_1 \rangle Cj_{R_1, R_2} \langle \psi_2, R_2 \rangle = \langle \psi', R_1 \cup R_2 \rangle, \text{ where } \langle \psi_1, R_1 \rangle \in \Psi(R_1),$$

$$\langle \psi_2, R_2 \rangle \in \Psi(R_2).$$

The basis of the multiset ψ' is defined by follow: $\Theta(\psi') = \{s \mid \exists s_1 \exists s_2 (s_1 \in \Theta(\psi_1) \wedge s_2 \in \Theta(\psi_2) \wedge s = s_1 \cup s_2)\}$. The number of duplicates is given by the following formula: $Occ(s, \psi') = Occ(s_1, \psi_1) \cdot Occ(s_2, \psi_2)$, where $s \in \Theta(\psi')$ and $s = s_1 \cup s_2$.

Definition 11. The Inner Natural Join of table on scheme R_1 and table on scheme R_2 is a binary parametric operation written as \otimes_{R_1, R_2} , whose value is the table on scheme $R_1 \cup R_2$ consisting of all the unions of compatible tuples of input tables. Hence, $\otimes_{R_1, R_2} : \Psi(R_1) \times \Psi(R_2) \rightarrow \Psi(R_1 \cup R_2)$, $\langle \psi_1, R_1 \rangle \otimes_{R_1, R_2} \langle \psi_2, R_2 \rangle = \langle \psi', R_1 \cup R_2 \rangle$, where $\psi_1 \in \Psi(R_1)$, $\psi_2 \in \Psi(R_2)$.

In other words, each tuple of ψ_1 is paired with each tuple of ψ_2 , regardless of whether it is a duplicate or not. The basis of the multiset ψ' is defined by follow: $\Theta(\psi') = \{s \mid \exists s_1 \exists s_2 (s_1 \in \Theta(\psi_1) \wedge s_2 \in \Theta(\psi_2) \wedge s_1 \approx s_2 \wedge s = s_1 \cup s_2)\}$. The number of duplicates is given by the following formula: $Occ(s_1 \cup s_2, \psi') = Occ(s_1, \psi_1) \cdot Occ(s_2, \psi_2)$, where $s' \in \Theta(\psi')$ and $s' = s_1 \cup s_2$. The relation \approx is a binary relation of compatibility of tuples $s_1 \approx s_2 \stackrel{dif}{\iff} s_1 \mid R = s_2 \mid R$ and $s_i \mid R$ is the restrictions of tuple s_i on the scheme R [5].

The Inner Join using A_1, \dots, A_n of table on scheme R_1 and table on scheme R_2 , moreover $R_1 \cap R_2 = \{A_1, \dots, A_n\}$, is a binary parametric operation of the form

$$\otimes_{A_1, \dots, A_n, R_1, R_2} : \Psi(R_1) \times \Psi(R_2) \rightarrow \Psi(R_1 \cup R_2),$$

$$\langle \psi_1, R_1 \rangle \otimes_{A_1, \dots, A_n, R_1, R_2} \langle \psi_2, R_2 \rangle = \langle \psi', R_1 \cup R_2 \rangle, \text{ where}$$

$$\langle \psi_1, R_1 \rangle \in \Psi(R_1), \langle \psi_2, R_2 \rangle \in \Psi(R_2).$$

Moreover all A_1, \dots, A_n are pairwise different, $n \geq 1$, and $R_1 \cap R_2 = \{A_1, \dots, A_n\}$. If input tables have also other general attributes which differ from A_1, \dots, A_n , before join they needs to be renamed.

The basis of the multiset ψ' is defined by follow: $\Theta(\psi') = \{s \mid \exists s_1 \exists s_2 (s_1 \in \Theta(\psi_1) \wedge s_2 \in \Theta(\psi_2) \wedge$

$\bigwedge_{i=1}^n s_i(A_i) = s_2(A_i) \wedge s = s_1 \cup s_2$ }. The number of duplicates is given by the following formula: $Occ(s, \psi') = Occ(s_1, \psi_1) \cdot Occ(s_2, \psi_2)$, where $s \in \Theta(\psi')$ and $s = s_1 \cup s_2$.

Let $p: S \times S \rightarrow \{true, false\}$ be a partial binary predicate on the set of all tuples S such that $\forall s_1 \forall s_2 ((s_1, s_2) \in \text{dom } p \wedge p(s_1, s_2) = true \Rightarrow s_1 \approx s_2)$.

Definition 12. The Inner Join on predicate p of table on scheme R_1 and table on scheme R_2 is a binary partial parametric operation of the form $\otimes_{p, R_1, R_2}: \Psi(R_1) \times \Psi(R_2) \rightarrow \Psi(R_1 \cup R_2)$, $\langle \psi_1, R_1 \rangle \otimes_{p, R_1, R_2} \langle \psi_2, R_2 \rangle = \langle \psi', R_1 \cup R_2 \rangle$.

The range of definition of this operation is $\text{dom } \otimes_{p, R_1, R_2} = \{ \langle \psi_1, R_1 \rangle, \langle \psi_2, R_2 \rangle \mid \Theta(\psi_1) \times \Theta(\psi_2) \subseteq \text{dom } p \}$. The basis of the multiset ψ' is defined by follow: $\Theta(\psi') = \{ s \mid \exists s_1 \exists s_2 (s_1 \in \Theta(\psi_1) \wedge s_2 \in \Theta(\psi_2) \wedge p(s_1, s_2) \approx true \wedge s = s_1 \cup s_2) \}$ and \approx is a generalized equality (strong Kleene's equality) [7]. The number of duplicates is given by the following formula: $Occ(s_1 \cup s_2, \psi') = Occ(s_1, \psi_1) \cdot Occ(s_2, \psi_2)$, where $s' \in \Theta(\psi')$ and $s' = s_1 \cup s_2$.

Let's note the following obvious fact. The join \otimes_{R_1, R_2} is extension of another arbitrary inner join operation in the following sense:

$$\begin{aligned} \langle t_1, R_1 \rangle Cj_{R_1, R_2} \langle t_2, R_2 \rangle &= \langle t_1, R_1 \rangle \otimes_{R_1, R_2} \langle t_2, R_2 \rangle, \\ \langle t_1, R_1 \rangle \otimes_{A_1, \dots, A_n, R_1, R_2} \langle t_2, R_2 \rangle &= \langle t_1, R_1 \rangle \otimes_{R_1, R_2} \langle t_2, R_2 \rangle, \\ \left(\langle t_1, R_1 \rangle \otimes_{p, R_1, R_2} \langle t_2, R_2 \rangle \right)_1 &\subseteq \left(\langle t_1, R_1 \rangle \otimes_{R_1, R_2} \langle t_2, R_2 \rangle \right)_1^a. \end{aligned}$$

The values of these operations in the left parts of these two equalities and inclusion must be defined.

Definition 13. The Semi-join of table on scheme R_1 and table on scheme R_2 is a binary parametric operation written as \bowtie_{R_1, R_2} , whose value is the table on scheme R_1 containing tuples of the first table which are included in the inner natural join of input tables. Thus, $\bowtie_{R_1, R_2}: \Psi(R_1) \times \Psi(R_2) \rightarrow \Psi(R_1)$, $\langle \psi_1, R_1 \rangle \bowtie_{R_1, R_2} \langle \psi_2, R_2 \rangle = \langle \psi', R_1 \rangle$, where $\langle \psi_1, R_1 \rangle \in \Psi(R_1)$, $\langle \psi_2, R_2 \rangle \in \Psi(R_2)$. The basis of the multiset ψ' is defined by follow: $\Theta(\psi') = \{ s_1 \mid s_1 \in \Theta(\psi_1) \wedge \exists s_2 (s_2 \in \Theta(\psi_2) \wedge s_1 \approx s_2) \}$.

^a $\left(\langle t, R \rangle \right)_1$ is the first component of the pair $\langle t, R \rangle$, i.e., is the set t .

The number of duplicates is given by the following formula: $Occ(s, \psi') = Occ(s, \psi_1)$, where $s \in \Theta(\psi')$.

B. Outer Join

We can lose information when using the inner join operations because the tuples which are not compatible will not be represented in the output table. The outer join operations use when it is necessary to consider the tuples of input tables which didn't get to result of the inner join operations.

Let $NULL$ be a special element of the universal domain D . $NULL$ used to denote absent values in the output table. Let $s_{R, NULL}$ be a constant tuple on scheme R , i.e. $s_{R, NULL}: R \rightarrow \{NULL\}$.

There is one logical scheme for definition of the outer join operations [5].

Let $\varphi: \Psi(R_1) \times \Psi(R_2) \rightarrow \Psi(R_1 \cup R_2)$ be some partial binary operation on the set of all tables and $(\varphi(\langle \psi_1, R_1 \rangle, \langle \psi_2, R_2 \rangle))_1 \subseteq \left(\langle \psi_1, R_1 \rangle \otimes_{R_1, R_2} \langle \psi_2, R_2 \rangle \right)_1$ for all tables $\langle \langle \psi_1, R_1 \rangle, \langle \psi_2, R_2 \rangle \rangle \in \text{dom } \varphi$.

Let's notice that the operations Cj_{R_1, R_2} , \otimes_{R_1, R_2} , $\otimes_{A_1, \dots, A_n, R_1, R_2}$, \otimes_{p, R_1, R_2} are such.

We fix two tables $\langle \psi_1, R_1 \rangle, \langle \psi_2, R_2 \rangle$ from range of definition of the operation φ .

Then the table $\langle \psi_1, R_1 \rangle$ takes the following form $\langle \psi_1, R_1 \rangle = \left\langle \psi_1 \cap_{\varphi} \psi_2, R_1 \right\rangle \cup_{All}^R \left\langle \psi_1 -_{\varphi} \psi_2, R_1 \right\rangle$.

Consider the table $\left\langle \psi_1 \cap_{\varphi} \psi_2, R_1 \right\rangle = \langle \psi', R_1 \rangle$. The basis of the multiset ψ' is defined by follow: $\Theta(\psi') = \{ s_1 \mid s_1 \in \Theta(\psi_1) \wedge \exists s_2 (s_2 \in \Theta(\psi_2) \wedge s_1 \cup s_2 \in \Theta(\varphi(\langle \psi_1, R_1 \rangle, \langle \psi_2, R_2 \rangle))) \}$. The number of duplicates is given by the following formula: $Occ(s_1, \psi') = Occ(s_1, \psi_1)$, where $s_1 \in \Theta(\psi')$.

Consider the table $\left\langle \psi_1 -_{\varphi} \psi_2, R_1 \right\rangle = \langle \psi'', R_1 \rangle$. The basis of the multiset ψ'' is defined by follow: $\Theta(\psi'') = \{ s_1 \mid s_1 \in \Theta(\psi_1) \wedge \forall s_2 (s_2 \in \Theta(\psi_2) \Rightarrow s_1 \cup s_2 \notin \Theta(\varphi(\langle \psi_1, R_1 \rangle, \langle \psi_2, R_2 \rangle))) \}$. The number of duplicates is given by the following formula: $Occ(s_1, \psi'') = Occ(s_1, \psi_1)$, where $s_1 \in \Theta(\psi'')$.

In other words, the tuples of the table $\langle \psi_1 \cap_{\varphi} \psi_2, R_1 \rangle$ are used in formation of result of the join operation, and tuples of the table $\langle \psi_1 -_{\varphi} \psi_2, R_1 \rangle$ are not used.

We obtain a representation of the table $\langle \psi_2, R_2 \rangle$ replacing the roles of the tables $\langle \psi_1, R_1 \rangle$ and $\langle \psi_2, R_2 \rangle$ in the presentation of the table $\langle \psi_1, R_1 \rangle$.

Let's notice that if the operation φ coincides with the operation \otimes_{R_1, R_2} then the table $\langle \psi_1 \cap_{\varphi} \psi_2, R_1 \rangle$ is the semi-join of the tables $\langle \psi_1, R_1 \rangle$ and $\langle \psi_2, R_2 \rangle$, i.e. $\langle \psi_1 \cap_{\varphi} \psi_2, R_1 \rangle = \langle \psi_1, R_1 \rangle \ltimes_{R_1, R_2} \langle \psi_2, R_2 \rangle$.

There are four kinds of the outer joins operations which are induced of the inner join operation φ : outer left join, outer right join, outer full join and union join. Let's define them.

Consider the following inner natural joins $\langle \psi_1 -_{\varphi} \psi_2, R_1 \rangle \otimes_{R_1, R_2 \setminus R_1} \langle \{s_{R_2 \setminus R_1, NULL}\}, R_2 \setminus R_1 \rangle = \langle \psi', R_1 \cup R_2 \rangle$, where $\Theta(\psi') = \{s_1 \cup s_{R_2 \setminus R_1, NULL} \mid s_1 \in \Theta(\psi_1 -_{\varphi} \psi_2)\}$, $Occ(s', \psi') = Occ(s_1, \psi_1 -_{\varphi} \psi_2)$, $s' \in \Theta(\psi')$, $s' = s_1 \cup s_{R_2 \setminus R_1, NULL}$ and $\langle \psi_2 -_{\varphi} \psi_1, R_2 \rangle \otimes_{R_2, R_1 \setminus R_2} \langle \{s_{R_1 \setminus R_2, NULL}\}, R_1 \setminus R_2 \rangle = \langle \psi'', R_1 \cup R_2 \rangle$, where $\Theta(\psi'') = \{s_2 \cup s_{R_1 \setminus R_2, NULL} \mid s_2 \in \Theta(\psi_2 -_{\varphi} \psi_1)\}$, $Occ(s'', \psi'') = Occ(s_2, \psi_2 -_{\varphi} \psi_1)$, $s'' \in \Theta(\psi'')$, $s'' = s_{R_1 \setminus R_2, NULL} \cup s_2$.

Definition 14. The Outer Left Join operation is a partial binary operation of the form $\varphi_l : \Psi(R_1) \times \Psi(R_2) \xrightarrow{\sim} \Psi(R_1 \cup R_2)$, where $\text{dom } \varphi_l = \text{dom } \varphi$ and $\varphi_l(\langle \psi_1, R_1 \rangle, \langle \psi_2, R_2 \rangle) = \varphi(\langle \psi_1, R_1 \rangle, \langle \psi_2, R_2 \rangle) \cup_{All}^{R_1 \cup R_2} \langle \psi_1 -_{\varphi} \psi_2, R_1 \rangle \otimes_{R_1, R_2 \setminus R_1} \langle \{s_{R_2 \setminus R_1, NULL}\}, R_2 \setminus R_1 \rangle$.

Definition 15. The Outer Right Join operation is a partial binary operation of the form $\varphi_r : \Psi(R_1) \times \Psi(R_2) \xrightarrow{\sim} \Psi(R_1 \cup R_2)$, where $\text{dom } \varphi_r = \text{dom } \varphi$ and $\varphi_r(\langle \psi_1, R_1 \rangle, \langle \psi_2, R_2 \rangle) = \varphi(\langle \psi_1, R_1 \rangle, \langle \psi_2, R_2 \rangle) \cup_{All}^{R_1 \cup R_2} \langle \psi_2 -_{\varphi} \psi_1, R_2 \rangle \otimes_{R_2, R_1 \setminus R_2} \langle \{s_{R_1 \setminus R_2, NULL}\}, R_1 \setminus R_2 \rangle$.

Definition 16. The Outer Full Join operation is a partial binary operation of the form $\varphi_f : \Psi(R_1) \times \Psi(R_2) \xrightarrow{\sim} \Psi(R_1 \cup R_2)$, where $\text{dom } \varphi_f = \text{dom } \varphi$

and $\varphi_f(\langle \psi_1, R_1 \rangle, \langle \psi_2, R_2 \rangle) = \varphi(\langle \psi_1, R_1 \rangle, \langle \psi_2, R_2 \rangle) \cup_{All}^{R_1 \cup R_2} \langle \psi_1 -_{\varphi} \psi_2, R_1 \rangle \otimes_{R_1, R_2 \setminus R_1} \langle \{s_{R_2 \setminus R_1, NULL}\}, R_2 \setminus R_1 \rangle \cup_{All}^{R_1 \cup R_2} \langle \psi_2 -_{\varphi} \psi_1, R_2 \rangle \otimes_{R_2, R_1 \setminus R_2} \langle \{s_{R_1 \setminus R_2, NULL}\}, R_1 \setminus R_2 \rangle$.

Definition 17. The Outer Union Join operation is a partial binary operation of the form $\varphi_U : \Psi(R_1) \times \Psi(R_2) \xrightarrow{\sim} \Psi(R_1 \cup R_2)$, where $\text{dom } \varphi_U = \text{dom } \varphi$ and $\varphi_U(\langle \psi_1, R_1 \rangle, \langle \psi_2, R_2 \rangle) =$

$$\langle \psi_1 -_{\varphi} \psi_2, R_1 \rangle \otimes_{R_1, R_2 \setminus R_1} \langle \{s_{R_2 \setminus R_1, NULL}\}, R_2 \setminus R_1 \rangle \cup_{All}^{R_1 \cup R_2} \langle \psi_2 -_{\varphi} \psi_1, R_2 \rangle \otimes_{R_2, R_1 \setminus R_2} \langle \{s_{R_1 \setminus R_2, NULL}\}, R_1 \setminus R_2 \rangle$$

C. Aggregate Operations

The five types of aggregate operations discussed in this article are SUM, AVERAGE, MAXIMUM, MINIMUM, COUNT. The aggregate operations transform a finite table into a table with single tuple and single attribute.

Consider the table $\langle \psi, R \rangle \in \Psi(R)$, where ψ is a finite multiset and $A \in R$. Let α_A be a multiset of column with attribute A of table $\langle \psi, R \rangle$ which contains all elements including duplicates.

Then $\Theta(\alpha_A) = \{d \mid \exists s (s \in \Theta(\psi) \wedge \langle A, d \rangle \in s)\} = \{d \mid \langle \langle A, d \rangle \rangle \in \Theta(\pi_{\{A\}, R}(\langle \psi, R \rangle))\}$ is an analogue of active domain of the attribute A [5]. The number of duplicates of element $d \in \Theta(\alpha_A)$ is given by the following formula: $\alpha_A(d) = Occ(\langle \langle A, d \rangle \rangle, (\pi_{\{A\}, R}(\langle \psi, R \rangle))_1) = \sum_{\substack{s \in \Theta(\psi) \\ s(A)=d}} Occ(s, \psi)$.

Let $2_m^{D'} = \{\alpha \mid \Theta(\alpha_A) \in 2^{D'}\}$ be a family of all multisets, bases of which are the finite subsets of the set D' . Here $D' \subseteq D$ is a subset of the universal domain.

Let Num is a numerical subset of the universal domain D that is closed under addition. Extend the set Num by the special element $NULL$. We will not extend the operation of addition to the case where at least one of the arguments is $NULL$.

Let's define the aggregate operations. At first the five aggregate functions – count, sum, average, maximum, minimum – are defined on a finite multiset and then these functions are transferred to the tables.

Definition 18. The aggregate operation $Sum_{A,R}$ by the attribute A of the finite table on scheme R , $A \in R$, is a unary parametric operation of the form $Sum_{A,R} : \Psi(R) \rightarrow \Psi(\{A\})$,

$Sum_{A,R}(\langle \psi, R \rangle) = \left\langle \left\{ \left\{ \left\langle A, Sum(\alpha_A) \right\rangle \right\}^1 \right\} \right\rangle^b$, where $\langle \psi, R \rangle \in \Psi(R)$. The $Sum(\alpha_A)$ function is applied to a column with attribute A in the table $\langle \psi, R \rangle$, the result obtained is the sum of every value occurrence in α_A . In addition, $NULL$ values don't undertake in attention and it is assumed that the column contains only data of numeric type.

Thus, $Sum: 2_m^{Num} \rightarrow Num$,

$$Sum(\alpha_A) = \begin{cases} NULL & \text{if } \Theta(\alpha_A) = \emptyset; \\ NULL & \text{if } \Theta(\alpha_A) = \{NULL\}; \\ \sum_{d \in \Theta(\alpha_A) \setminus \{NULL\}} d\alpha_A(d) & \text{if } \Theta(\alpha_A) \setminus \{NULL\} \neq \emptyset. \end{cases}$$

So, we have $Sum(\{NULL^n\}) = NULL$,

$Sum(\{d_1^{n_1}, \dots, d_k^{n_k}\}) = \sum_{i=1}^k d_i n_i$ if all elements $d_i, i = \overline{1, k}$, differ from $NULL$.

In the case of the empty table $\langle \psi_\emptyset, R \rangle$ we have

$$Sum_{A,R}(\langle \psi_\emptyset, R \rangle) = \left\langle \left\{ \left\langle A, NULL \right\rangle \right\}^1 \right\rangle \{A\}, \text{ here } \psi_\emptyset = \emptyset_m.$$

Example 1. Let $\langle \psi, R \rangle$ be the table of Fig 1. Then

$$Sum_{A,R}(\langle \psi, R \rangle) = \left\langle \left\{ \left\langle A, 8 \right\rangle \right\}^1 \right\rangle \{A\},$$

$$Sum_{B,R}(\langle \psi, R \rangle) = \left\langle \left\{ \left\langle B, 6 \right\rangle \right\}^1 \right\rangle \{B\},$$

$Sum_{C,R}(\langle \psi, R \rangle) = \left\langle \left\{ \left\langle C, 6 \right\rangle \right\}^1 \right\rangle \{C\}$. In Fig. 2, Fig. 3 and Fig. 4 reactively, we see the tables $Sum_{A,R}(\langle \psi, R \rangle)$, $Sum_{B,R}(\langle \psi, R \rangle)$, $Sum_{C,R}(\langle \psi, R \rangle)$.

A	B	C
NULL	0	3
2	1	1
2	1	1
2	1	1
2	3	NULL

Fig. 1. Table $\langle \psi, R \rangle$

A
8

Fig. 2. Table $Sum_{A,R}(\langle \psi, R \rangle)$

B

^b The top index 1 specifies that the table include the tuple $\{\langle A, Sum(\alpha_A) \rangle\}$ only once, i.e. $\{\{\langle A, Sum(\alpha_A) \rangle\}^1\}$ is $\{1\}$ -multiset.

6

Fig. 3. Table $Sum_{B,R}(\langle \psi, R \rangle)$

C
6

Fig. 4. Table $Sum_{C,R}(\langle \psi, R \rangle)$

Let \leq be a linear order on the universal domain D .

Definition 19. The aggregate operation $Min_{A,R}$ by the attribute A of the finite table on scheme $R, A \in R$, is a unary parametric operation of the form $Min_{A,R}: \Psi(R) \rightarrow \Psi(\{A\})$,

$$Min_{A,R}(\langle \psi, R \rangle) = \left\langle \left\{ \left\langle A, Min(\alpha_A) \right\rangle \right\}^1 \right\rangle \{A\},$$

where $\langle \psi, R \rangle \in \Psi(R)$. The $Min(\alpha_A)$ function is applied to a column with attribute A in the table $\langle \psi, R \rangle$, the result obtained is the minimum value among values of α_A . In addition, $NULL$ values don't undertake in attention.

Thus, $Min: 2_m^D \rightarrow D$,

$$Min(\alpha_A) = \begin{cases} NULL & \text{if } \Theta(\alpha_A) = \emptyset; \\ NULL & \text{if } \Theta(\alpha_A) = \{NULL\}; \\ \min\{d \mid d \in \Theta(\alpha_A) \setminus \{NULL\}\} & \text{if } \Theta(\alpha_A) \setminus \{NULL\} \neq \emptyset. \end{cases}$$

We have $Min(\emptyset_m) = NULL, Min(\{NULL^n\}) = NULL, Min(\{d_1^{n_1}, \dots, d_k^{n_k}\}) = \min\{d_1, \dots, d_k\}$ if all elements $d_i, i = \overline{1, k}$, differ from $NULL$.

In the case of the empty table $\langle \psi_\emptyset, R \rangle$ we have

$$Min_{A,R}(\langle \psi_\emptyset, R \rangle) = \left\langle \left\{ \left\langle A, NULL \right\rangle \right\}^1 \right\rangle \{A\}, \text{ here } \psi_\emptyset = \emptyset_m.$$

Example 2. Let $\langle \psi, R \rangle$ be the table of Fig 1. Then

$$Min_{A,R}(\langle \psi, R \rangle) = \left\langle \left\{ \left\langle A, 2 \right\rangle \right\}^1 \right\rangle \{A\},$$

$$Min_{B,R}(\langle \psi, R \rangle) = \left\langle \left\{ \left\langle B, 0 \right\rangle \right\}^1 \right\rangle \{B\},$$

$$Min_{C,R}(\langle \psi, R \rangle) = \left\langle \left\{ \left\langle C, 1 \right\rangle \right\}^1 \right\rangle \{C\}.$$

Definition 20. The aggregate operation $Max_{A,R}$ by the attribute A of the finite table on scheme $R, A \in R$, is a unary parametric operation of the form $Max_{A,R}: \Psi(R) \rightarrow \Psi(\{A\})$,

$$Max_{A,R}(\langle \psi, R \rangle) = \left\langle \left\{ \left\langle A, Max(\alpha_A) \right\rangle \right\}^1 \right\rangle \{A\},$$

where $\langle \psi, R \rangle \in \Psi(R)$. The $Max(\alpha_A)$ function is applied to a column with attribute A in the table $\langle \psi, R \rangle$, the result obtained is the maximum value among values of α_A . In addition, $NULL$ values don't undertake in attention.

Thus, $Max: 2_m^D \rightarrow D$,

$$Max(\alpha_A) = \begin{cases} NULL & \text{if } \Theta(\alpha_A) = \emptyset; \\ NULL & \text{if } \Theta(\alpha_A) = \{NULL\}; \\ \max\{d \mid d \in \Theta(\alpha_A) \setminus \{NULL\}\} & \text{if } \Theta(\alpha_A) \setminus \{NULL\} \neq \emptyset. \end{cases}$$

We have $Max(\emptyset_m) = NULL$, $Max(\{NULL^n\}) = NULL$,
 $Max(\{d_1^{n_1}, \dots, d_k^{n_k}\}) = \max\{d_1, \dots, d_k\}$ if all elements d_i ,
 $i = \overline{1, k}$, differ from $NULL$.

In the case of the empty table $\langle \psi_\emptyset, R \rangle$ we have
 $Max_{A,R}(\langle \psi_\emptyset, R \rangle) = \langle \{A, NULL\}^1 \rangle, \{A\}$, here $\psi_\emptyset = \emptyset_m$.

Example 3. Let $\langle \psi, R \rangle$ be the table of Fig 1. Then
 $Max_{A,R}(\langle \psi, R \rangle) = \langle \{A, 2\}^1 \rangle, \{A\}$,
 $Max_{B,R}(\langle \psi, R \rangle) = \langle \{B, 3\}^1 \rangle, \{B\}$,
 $Max_{C,R}(\langle \psi, R \rangle) = \langle \{C, 3\}^1 \rangle, \{C\}$.

Definition 21. The aggregate operation $Count_{A,R}$ by the
 attribute A of the finite table on scheme R , $A \in R$, is a
 unary parametric operation of the form
 $Count_{A,R}: \Psi(R) \rightarrow \Psi(\{A\})$,

$Count_{A,R}(\langle \psi, R \rangle) = \langle \{A, Count(\alpha_A)\}^1 \rangle, \{A\}$, where
 $\langle \psi, R \rangle \in \Psi(R)$. The $Count(\alpha_A)$ function is applied to a
 column with attribute A in the table $\langle \psi, R \rangle$, the result
 obtained is the count of all values of α_A which differ from
 $NULL$.

Thus, $Count: 2_m^D \rightarrow N$, $Count(\alpha_A) = \sum_{d \in \Theta(\alpha_A) \setminus \{NULL\}} \alpha_A(d)$. Put
 by definition that the sum of an empty set of elements is equal
 to zero.

So, we have $Count(\emptyset_m) = 0$, $Count(\{NULL^n\}) = 0$,
 $Count(\{d_1^{n_1}, \dots, d_k^{n_k}\}) = n_1 + \dots + n_k$ if all elements d_i , $i = \overline{1, k}$,
 differ from $NULL$.

In the case of the empty table $\langle \psi_\emptyset, R \rangle$ we have
 $Count_{A,R}(\langle \psi_\emptyset, R \rangle) = \langle \{A, 0\}^1 \rangle, \{A\}$, here $\psi_\emptyset = \emptyset_m$.

Example 4. Let $\langle \psi, R \rangle$ be the table of Fig 1. Then
 $Count_{A,R}(\langle \psi, R \rangle) = \langle \{A, 4\}^1 \rangle, \{A\}$,
 $Count_{B,R}(\langle \psi, R \rangle) = \langle \{B, 5\}^1 \rangle, \{B\}$,
 $Count_{C,R}(\langle \psi, R \rangle) = \langle \{C, 4\}^1 \rangle, \{C\}$.

We assume that a numerical subset Num of the universal
 domain D is closed under the (partial operation) division
 operation $/: Num \times Num \rightsquigarrow Num$. We will determine the
 division operation so that when the first argument is equal to
 $NULL$ the function accepts value $NULL$.

Definition 22. The aggregate operation $Avg_{A,R}$ by the
 attribute A of the finite table on scheme R , $A \in R$, is a
 unary parametric operation of the form
 $Avg_{A,R}: \Psi(R) \rightarrow \Psi(\{A\})$,

$Avg_{A,R}(\langle \psi, R \rangle) = \langle \{A, Avg(\alpha_A)\}^1 \rangle, \{A\}$, where
 $\langle \psi, R \rangle \in \Psi(R)$. The $Avg(\alpha_A)$ function is applied to a column
 with attribute A in the table $\langle \psi, R \rangle$, the result obtained is the
 arithmetic mean of values in α_A which differ from $NULL$.

$$\text{Thus, } Avg: 2_m^{Num} \rightarrow Num \text{ and } Avg(\alpha_A) = \frac{Sum(\alpha_A)}{Count(\alpha_A)}.$$

$$\text{We have } Avg(\emptyset_m) = \frac{Sum(\emptyset_m)}{Count(\emptyset_m)} = \frac{NULL}{0} = NULL,$$

$$Avg(\{NULL^n\}) = \frac{Sum(\{NULL^n\})}{Count(\{NULL^n\})} = \frac{NULL}{0} = NULL,$$

$$Avg(\{d_1^{n_1}, \dots, d_k^{n_k}\}) = \frac{Sum(\{d_1^{n_1}, \dots, d_k^{n_k}\})}{Count(\{d_1^{n_1}, \dots, d_k^{n_k}\})} = \frac{\sum_{i=1}^k d_i n_i}{(n_1 + \dots + n_k)} \text{ if}$$

all elements d_i , $i = \overline{1, k}$, differ from $NULL$.

In the case of the empty table $\langle \psi_\emptyset, R \rangle$ we have
 $Avg_{A,R}(\langle \psi_\emptyset, R \rangle) = \langle \{A, NULL\}^1 \rangle, \{A\}$, here $\psi_\emptyset = \emptyset_m$.

Example 5. Let $\langle \psi, R \rangle$ be the table of Fig 1. Then
 $Avg_{A,R}(\langle \psi, R \rangle) = \langle \{A, 2\}^1 \rangle, \{A\}$,
 $Avg_{B,R}(\langle \psi, R \rangle) = \langle \{B, \frac{6}{5}\}^1 \rangle, \{B\}$,
 $Avg_{C,R}(\langle \psi, R \rangle) = \langle \{C, \frac{6}{4}\}^1 \rangle, \{C\}$.

Definition 23. The aggregate operation $Count_{A,R}(\ast)$ by the
 attribute A of the finite table on scheme R , $A \in R$, is a
 unary parametric operation of the form
 $Count_{A,R}(\ast): \Psi(R) \rightarrow \Psi(\{A\})$,

$Count_{A,R}(\ast)(\langle \psi, R \rangle) = \langle \{A, \|\psi\|\}^1 \rangle, \{A\}$, where
 $\langle \psi, R \rangle \in \Psi(R)$, and $\|\psi\|$ is the rank of the multiset ψ .

The operation $Count_{A,R}(*)$ finds the number of tuples in the table $\langle \psi, R \rangle$.

In the case of an empty table $\langle \psi_{\emptyset}, R \rangle$ we have $Count_{A,R}(*)(\langle \psi_{\emptyset}, R \rangle) = \langle \{ \langle A, \|\emptyset_m\| \rangle \}, \{A\} \rangle = \langle \{ \langle A, 0 \rangle \}, \{A\} \rangle$, here $\psi_{\emptyset} = \emptyset_m$.

Example 5. Let $\langle \psi, R \rangle$ be the table of Fig 1. Then

$$Count_{A,R}(*)(\langle \psi, R \rangle) = \langle \{ \langle A, 5 \rangle \}, \{A\} \rangle,$$

$$Count_{B,R}(*)(\langle \psi, R \rangle) = \langle \{ \langle B, 5 \rangle \}, \{B\} \rangle,$$

$$Count_{C,R}(*)(\langle \psi, R \rangle) = \langle \{ \langle C, 5 \rangle \}, \{C\} \rangle.$$

V. CONCLUSIONS

In this paper the multiset table algebra is considered. The signature of the multiset table algebra is filled up with new operations such as inner and outer join, semijoin and aggregate operations. The special element NULL is inserted in the universal domain for a define of outer operations.

It should also be noted that a parameter of aggregate operations is not necessarily only a single attribute; it also can be some function of the tuples.

- [1] E. F. Codd, "A Relational Model for Large Shad Data Banks," in *Communications of the ACM*, USA, New York, 1970, vol.13, No.6, pp. 377-387.
- [2] Paul W.P.J. Grefen and Rolf A. de By, "A Multi-Set Extended Relational Algebra. A Formal Approach to a Practical Issue," in *10th International Conference on Data Engineering*, Houston, TX, USA, 1994, pp. 80-88.
- [3] G. Lamperti, M. Melchiori, and M. Zanella, "On Multisets in Database Systems," in *Proceedings of the Workshop on Multiset Processing: Mathematical, Computer Science, and Molecular Computing Points of View*, London, UK, 2001, pp. 147-216.
- [4] H. Garcia-Molina, J. D. Ullman, J. Widom, "Algebraic and Logical Query Languages," *Database Systems: The Complete Book*, in 2th ed. New Jersey, Upper Saddle River, 2009, ch. 2, sec. 5, pp. 203-241.
- [5] V. Redko, J. Brona, D. Buy, S. Poliakov, "Compositional semantics of SQL," in *Relation Database: Relation Algebras and SQL-similar Languages*, Kyiv, Ukraine, 2001, ch. 3, sec. 3.4, pp. 151-180.
- [6] D. B. Buy, I. M. Glushko, "Extended of Table Algebra: Multiset Table Algebra," in *Modern scientific research and their practical application*, Ukraine, Odessa, 2013, vol.J11309, Article CID Number 261.
- [7] N. Cutland, "Prologue. Prerequisites and notation" in *Computability. An introduction to recursive function theory*. London, Cambridge University Press, 1980, sec. 2, pp. 2-4.

Temporal Data in Enterprise Database Systems

Dušan Petković

University of Applied Sciences
Rosenheim, 83024, Germany
petkovic@fh-rosenheim.de

Abstract—The time is generally a challenging task. All issues in relation to time can be better supported using temporal data models. More than two-dozen such models have been introduced in time period between 1988 and 1995. After this period, the work on temporal data has been not so dynamic. The last couple of years brought the new revival of the topic and the emergence of new data models. In this article we present the final release of temporal data model, which is published as the part of the latest SQL standard. After that we discuss the differences between the specification and one of its implementations. Finally, we conclude the article with discussion of properties of this data model and advocate for introduction of the PERIOD data type in one of the future versions of the specification.

Keywords— *temporal databases, valid time, transaction time, bitemporal, standardization, SQL standard*

I. INTRODUCTION

After many years of hard work, the ISO standardization committee for SQL has released the final specification for temporal data. This specification has its roots in several proposals, which came from different sources. One of the most important differences to the previous proposals is that the specification does not build an entirely new part, as planned before.

The story of temporal data specification for the SQL standard is very long and rich on twists. In the year 1994, the ANSI department working on the SQL standard had started the work on temporal data. (The American National Standards Institute - ANSI - is the organization that oversees the development of standards in the United States.) They made a proposal, which was based upon the work of Richard Snodgrass and his colleagues. Professor Snodgrass has previously published a specification of temporal language, which was an extension of the SQL standard at that time [6]. The language specification, together with other materials has been published later in a book [7]. The American proposal has not been accepted by the ISO committee due to several significant insufficiencies [1]. (The International Standard Organization, ISO, is the international counterpart to ANSI.) At the same time, the members of the English standardization committee made another proposal, which was based upon the work of Nikos Lorentzos [3].

As a reaction to the reject of the American proposal, the members of the ANSI committee did not agree with the proposal of their British colleagues, hence none of these specifications has been accepted at that time. For this reason, the next SQL standard, SQL:1999, [4] did not contain the specification of temporal data at all. In the following years, there were no attempts to solve this problem and to make a specification for temporal data. There was a deadlock between

members of the ISO committee, and hence, in the year 2001 the SQL standardization committee decided to abandon the work on temporal data.

In the year 2007, the members of SQL standardization committee started the work concerning system- versioned tables. At the same time they made the decision to add the already existing specification to SQL/Foundation. (The last SQL standard comprises nine parts, which are not consecutively numbered. The most important part is the second one, SQL/Foundation, which comprises the foundations of the language. For this reason, this part is the most voluminous of all existing parts.)

In the year 2010, the committee started the work concerning application-time period tables. The both extensions, application-time period tables and system-versioned tables, build the biggest part of the new specification for temporal data, which has been released in the SQL:2011 standard. Generally, the new specification inherits a lot of ideas from the both previous proposals, has however a significantly different syntax. (A list of all temporal extensions in SQL:2011 can be found in [2], while all non-temporal extensions are described in [9].)

This article presents the temporal data model proposed by the SQL standardization committee. We also discuss the differences between this specification and the implementation of temporal data in IBM DB2. The main contribution of this paper is to show deficiencies of the proposed model and to present the PERIOD data type, which should be considered in future specifications.

The rest of this article is organized in the following way: Section 2 deals with the syntax extensions in relation to application-time period tables. The new syntax of CREATE TABLE e.g. ALTER TABLE statement is given, as well as the syntax extensions in INSERT, DELETE and UPDATE

statements. Section 3 describes system-versioned tables. Section 4 discusses a specification of a new table form (called bitemporal table), where application-time period data as well as system-versioned data are combined. Section 5 shows the implementation of system-versioned, application-time period and bitemporal tables in IBM DB2. The last section gives conclusions and discusses future work.

II. APPLICATION-TIME PERIOD TABLES

An application-time period table is a table that contains a PERIOD clause with a user-defined name for time period. The SQL:2011 standard restricts such a table so that its rows are associated with one or more temporal periods. A typical example of such a problem is an insurance application, where it is necessary to keep track of insurance information (art of insurance, annual premium etc.) of a given customer that are in effect at any given point in time.

An application-time period table contains two additional columns, one to store the start time of a period associated with the row and one to store the end time. Values of both columns are set by the user. Additional syntax is provided for users to specify primary key/unique constraints to ensure that no two rows with the same key value have overlapping periods.

Note that application-time period tables are related to a temporal dimension called valid time. Valid time concerns the time when an event is true in the real world. For this reason, this form of time is independent of its storage in a database and can concern the past, present and future snapshots of the event. Using timestamps, it is possible to form different versions of an event. This is the central aspect of a temporal database realization.

A. Creating Application-Time Period Tables

When creating an application-time period table, two additional columns must be defined. The former stores start values and the latter the end values of the corresponding time period. The both columns must be NOT NULL and their data type can be either DATE or TIMESTAMP. The interval specified by the values of these columns is half open, meaning that it contains the value of the start column but not the value of the end column. The both columns are specified in the CREATE TABLE or ALTER TABLE statement, using the PERIOD clause. This clauses specify names of both columns explicitly and the implicit rule that start_date<end_date. (The name of this time period is specified by the user.) Example 1 shows the creation of an application-time period table.

Example 1

```
CREATE TABLE a_employees (emp_id VARCHAR(30) NOT NULL,
    dept_name VARCHAR (20) NOT NULL,dept_id VARCHAR(30),
    start DATE NOT NULL, end DATE NOT NULL,
```

```
PERIOD FOR emp_period (start, end), PRIMARY KEY (emp_id,
emp_period WITHOUT OVERLAPS),
    FOREIGN KEY (dept_id, PERIOD emp_period)
REFERENCES department (dept_id, PERIOD dept_period));
```

The example above shows two other extensions of the CREATE TABLE statement in relation to application-time period tables: The first one is the WITHOUT OVERLAPS clause. This clause forbids overlapping of time periods for the same value of non-temporal part of the primary key. Additionally, the specification of the PERIOD clause in the FOREIGN KEY option forbids the existence of a row in a referencing table whose time period is not contained in the time period of a corresponding referenced table.

B. Retrieving and Modifying Data from Application-Time Period Tables

The syntax of the INSERT statement for application-time period tables is identical to the syntax of the same statement for convenient tables. This means that the start and end time of the period has to be explicitly specified by the user. (The both values can be related to the past, present or future.) Example 2 shows the insertion of a row in the **a_employees** table, while Table 1 displays the table's content after insertion.

Example 2

```
INSERT INTO a_employees
(emp_id, dept_name, dept_id, start, end) VALUES
('e1', 'Marketing', 'd1', DATE'2010-01-15', DATE '2011-01-15');
```

Table 1: The content of the **a_employees** table

id	name	id	start	End
e1	Market.	d1	2010.1.15	2011.1.15

As we already stated, the WITH OVERLAPS clause forbids overlapping of time periods for the same value of the non-temporal part of the primary key. For this reason, the INSERT statement in Example 3 will produce an error.

Example 3

```
INSERT INTO a_employees
(emp_id, dept_name, dept_id, start, end) VALUES
('e1', 'Marketing', 'd1', DATE'2010-04-01', DATE '2010-12-31');
```

The insertion of a row into the **a_employees** table in Example 3 will not be executed, because of the existence of the **emp_period** integrity rule in the PRIMARY KEY clause in Example 1. The time period of the row in Example 3 ('2010-04-01', '2010-12-31') overlaps the time period of the inserted row ('2010-01-15', '2011-01-15').

The syntax of the UPDATE statement is extended with the FOR PORTION clause to support temporal data. This clause is used to specify the time period for which the

modification in the UPDATE statement is applied. Example 4 shows the use of this clause and Table 2 displays the content of the table after modification.

Example 4

```
UPDATE a_employees FOR PORTION OF emp_period
FROM DATE '2010-05-01' TO DATE '2010-08-01'
SET dept_id = 'd2' WHERE emp_id = 'e1';
```

Table 2: Content of a_employees after modification

Id	Dept	Start	End
e1	Marketing	10.1.15	10.05.1
e1	Marketing	10.5.1	10.08.01
e1	Marketing	10.8.1	11.01.15

The time period specified in the FOR PORTION clause in Example 4 divides the time period of the already inserted row in two parts. For this reason after the execution of the UPDATE statement the table will contain two new inserted rows and the previous row, with the modified time period. The DELETE statement can be used with its convenient syntax or with the same extension as the UPDATE statement. In Example 5, the FOR PORTION clause specifies the time period, for which the deletion is applied.

Example 5

```
DELETE FROM a_employees FOR PORTION OF emp_period
FROM DATE '2010-06-01' TO DATE '2011-01-01';
```

Table 3: The result of DELETE in Example 5

Emp	Dept	Start	End
e1	d1	2010.1.15	2010.5.01
e1	d2	2010.5.01	2010.6.01
e1	d1	2011.1.01	2011.1.15

The DELETE statement in Example 5 concerns the time periods of the second and third row in Table 2: For this reason, these two time periods will be “shortened” according to the specified period('2010-06-01', '2011-01-01'). Table 3 shows the content of the table after execution of the DELETE statement.

III. SYSTEM-VERSIONED TABLES

System-versioned tables are intended to solve real world problems, where the history of data modifications must be maintained. The structure of system-versioned tables is extended with two new columns that contains begin and the end of the specified time period. The values of these columns contain system times, which are updated each time the table content is modified.

System-versioned tables are related to a temporal dimension called transaction time. Transaction time

concerns the time the fact was present in the database as stored data. In other words, the transaction time of an event describes the times, where the event is stored in a database and presents the correct image of the modelled world. Timestamps of transaction time events are defined according to the schedule adopted by the operating system. Therefore, we can build the history of all such timestamps in relation to the past and current time, but not in relation to future. For this reason, system-versioned tables contain system times, which are updated each time the table content is modified.

A. Creating System-Versioned Tables

The names of the two new columns described above are specified by the user, but their values are inserted by the system. The syntax of the CREATE TABLE statement contains several new extensions, which can be seen in Example 6

Example 6

```
CREATE TABLE s_employees
(emp_name VARCHAR(50) NOT NULL,
dept_id VARCHAR(10), system_start
TIMESTAMP(12) GENERATED ALWAYS AS ROW START,
system_end TIMESTAMP(12)
GENERATED ALWAYS AS ROW END, PERIOD FOR
SYSTEM_TIME (system_start, system_end), PRIMARY KEY
(emp_name)) WITH SYSTEM VERSIONING;
```

The new clauses in CREATE TABLE statement in relation to system-versioned tables are:

- GENERATED ALWAYS AS ROW START
- GENERATED ALWAYS AS ROW END

The former clause specifies begin of the time period, while the latter defines the end of that period. Therefore, the column **system_start** in Example 6 stores values in relation to begin of the system time period and the column **system_end** the values of the end of it. The PERIOD FOR SYSTEM TIME clause contains the names of both columns and the given order of them implies the rule that the value of the first column must be always earlier than that of the second one. The last option, WITH SYSTEM VERSIONING, inserts implicitly the start time values to the corresponding values of the column, which build the primary key. The reason for this is that in a case of system-versioned tables, the values of a non-temporal column, which builds the primary key, are not unique, because several versions of such a column can exist. An instance of temporal entity must have a primary key composed of time-varying and non-time-varying attributes. Therefore, the values of activation start time are used as the part of the primary key.

The most important attitude of system-versioned tables is that old versions of an instance are preserved. In spite of it,

the current instance (one which contains the current time) and those with previous (historical) times are treated differently: The former is called current system row and only that one will be modified with update operations. The old versions of the same table are called historical rows and they are read-only. Note, that already specified constraints are valid only for the current row(s).

B. Retrieving and Modifying Data from System-Versioned Tables

The syntax of the INSERT statement for system-versioned tables is identical to the syntax of the same statement for convenient tables. The values of both system time columns are implicitly inserted by the system. If a new row is inserted in the system-versioned table, the current time is assigned to the column with the start time, while the highest possibly timestamp is inserted into the column with the end time. Example 7 uses the INSERT statement to add a new row to the s_employees table, while Table 4 displays the content of the table after insertion. Note that although the time-variant columns of the s_employees table are defined using the TIMESTAMP data type, only the DATE portion of them will be shown in results, because of simplicity. (The assumed current time for this example is 2012.08.01.)

Example 7

```
INSERT INTO s_employees (emp_name, dept_id)
VALUES ('Scott', 'd1');
```

Table 4: The content of s_employees after insertion

id	emp_name	sys_start	sys_end
d1	Scott	2012.8.1	9999.12.31

Concerning the UPDATE statement, the columns with the start and end time cannot be used explicitly in the SET clause of that statement. When a row of the system-versioned table is modified, the old version of that row is preserved, before the column values are modified and the current version is inserted. At the same time, the end time of the old version and the begin time of the new one will be set to the current (transaction) time. (The DELETE statement has the same semantics as the UPDATE statement.)

The modification of the s_employees table is shown in Example 8, while Table 5 displays the content of that table after update. Note that the “deleted” rows still belong to the content of the table. Only their activation end time will be set to the current time. (The current of execution the UPDATE statement time is 2012.08.08.)

Example 8

```
UPDATE s_employees
SET dept_id = 'd2'
WHERE emp_name = 'Scott';
```

Table 5: The content of the employee table after update

dept_id	name	sys_start	sys_end
d1	Scott	2012.08.01	2012.08.08
d2	Scott	2012.08.08	9999.12.31

The syntax of the SELECT statement in relation to system-versioned tables is the same as for the regular tables. The only difference is the necessity to retrieve old versions of rows. This can be done using the FOR SYSTEM_TIME clause. The meaning of this clause is to deliver rows, which satisfy the given condition. The resulting rows can be current or old version rows, depending on the condition.

There are four different forms of this clause:

- FOR SYSTEM_TIME AS OF CURRENT TIMESTAMP
- FOR SYSTEM_TIME AS OF <datetime value expression>
- FOR SYSTEM_TIME BETWEEN < date value expr 1> AND< date value expr 2>
- FOR SYSTEM_TIME FROM < date value expr 1> TO< date value expr 2>

The first form of the clause is the default value. This means that if a query includes any explicit form of the FOR SYSTEM_TIME clause, this form is implicitly assumed and the query returns the current rows as the result. The second form of the clause is used to retrieve rows of a table at a specified point in time. In contrast to the second form of the FOR SYSTEM_TIME clause, the third and the fourth form specify the condition as a time period. (The former defines a closed interval, while the latter specifies a half open interval.) Examples 9 and 10 show second and the third form of the FOR SYSTEM_TIME clause, respectively.

Example 9

```
SELECT dept_name
FROM s_employees FOR SYSTEM_TIME AS OF DATE
'2010-01-01' WHERE emp_name = 'Scott';
```

Example 10

```
SELECT dept_name
FROM s_employees FOR SYSTEM_TIME BETWEEN DATE
'2010-01-01' AND DATE '2012-01-01'
WHERE emp_name = 'Scott';
```

IV. BITEMPORAL TABLES

A bitemporal table comprises both an application-time period table as well as a system-versioned table. To understand why the “marriage” of both table forms is useful in the real world, let us take a look at an example. During their existence, departments of a firm can change their names. Typically, the

modification of a department name happens at a specific time, but that name is changed in the database not at the same time (usually later). In that case, the system-time period automatically records when a particular name is inserted into the database and the application time period records when the name was actually modified.

A. Creating Bitemporal Tables

When creating a bitemporal table, four additional temporal columns must be specified, two concerning system times and two in relation to application-time period. Example 11 shows the creation of such a table. (The CREATE TABLE statement in Example 11 is just a union of columns and clauses from Example 1 and Example 6.)

Example 11

```
CREATE TABLE bi_employees (emp_id VARCHAR(30) NOT NULL,  
dept_name VARCHAR (20) NOT NULL, dept_id VARCHAR(30),  
start DATE NOT NULL, end DATE NOT NULL,  
system_start DATE NOT NULL  
GENERATED ALWAYS AS ROW START,  
system_end DATE NOT NULL GENERATED ALWAYS AS ROW  
END, PERIOD FOR SYSTEM_TIME (system_start, system_end),  
PERIOD FOR emp_period (start, end),  
PRIMARY KEY (emp_id, emp_period WITHOUT OVERLAPS))  
WITH SYSTEM VERSIONING;
```

Because bitemporal tables combine properties of both forms of temporal tables, all DML statements (SELECT, INSERT, UPDATE and DELETE) can be used either for the application-time period, system-time period or the combination or both. In other words, there are no syntactic extensions which are specific for bitemporal tables.

V. IMPLEMENTATION OF TEMPORAL DATA IN IBM DB2

Several vendors of RDBMSs have already implemented temporal data. Some of them used the prerelease of the specification for implementation, while others used proprietary syntax and semantics. At this moment, there is only one vendor, who implemented the specification of temporal data from the SQL standard: IBM DB2. For this reason, we will describe the support of temporal data in IBM DB2 in this section.

A. Tables with Business Time

The implementation of application-time period tables in DB2 is similar to the corresponding specification in SQL:2011: The main difference is in the terminology: In DB2, such tables are called tables with business time [5].

1) Creating Tables with Business Time

The syntax of the CREATE TABLE statement for creation of tables with business time is slightly different to the corresponding syntax for creation of application-time period tables. Example 12 shows how the CREATE TABLE statement can be used to create tables with business time.

Example 12

```
CREATE TABLE a_employees (emp_id VARCHAR(30) NOT NULL,  
dept_name VARCHAR (20) NOT NULL, dept_id VARCHAR(30),  
start DATE NOT NULL, end DATE NOT NULL,  
PERIOD BUSINESS_TIME (start, end),  
PRIMARY KEY (emp_id, BUSINESS_TIME WITHOUT  
OVERLAPS));
```

The only difference to the corresponding SQL:2011 specification is that IBM DB2 does not allow users to define names for the specified time period (in the PERIOD clause). This is replaced by the BUSINESS_TIME reserved keyword.

2) Modifying Data from Tables with Business Time

The syntax of the INSERT statement in DB2 is identical to the syntax of the same statement in the SQL:2011 specification, while the syntax of the UPDATE and DELETE statements in DB2 is slightly different: The FOR PORTION clause contains the BUSINESS_TIME reserved word, instead of the user-defined name. (This is an implication from the definition of the PERIOD clause in the CREATE TABLE statement in DB2.) Example 13 shows the use of the UPDATE statement to modify tables with business time, while deletion of rows is given in Example 14. (These two examples correspond to Examples 4 and 5. For this reason, the result of these two statements is given in Table 2 and Table 3, respectively.)

Example 13

```
UPDATE a_employees FOR PORTION OF BUSINESS_TIME  
FROM DATE '2010-05-01' TO DATE '2010-08-01'  
SET dept_id = 'd2' WHERE emp_id = 'e1';
```

Example 14

```
DELETE FROM a_employees  
FOR PORTION OF BUSINESS_TIME  
FROM DATE '2010-06-01' TO DATE '2011-01-01';
```

B. Tables with System Times

The semantics of system-versioned tables in DB2 is different than the semantics of the corresponding specification in SQL:2011. Instead of one system-versioning table, DB2 supports two tables, one to store current rows and one to store old versions of them. Besides that, the terminology is different: System-versioning tables are called tables with system time.

There are three steps in defining tables with system time:

- Create the base table (for current rows)
- Create the versioning table (for old versions of rows)

- Alter the base table to enable versioning and identify the versioning table

1) Creating Tables with System Time

The syntax of the CREATE TABLE statement for creating base tables with business time is almost identical to the corresponding syntax for creation of system-versioning tables. Example 15 shows how the CREATE TABLE statement can be used to create a base table with system time.

Example 15

```
CREATE TABLE s_employees (emp_name VARCHAR(50) NOT
NULL, dept_id VARCHAR(10),
system_start TIMESTAMP(12) GENERATED ALWAYS AS ROW
BEGIN NOT NULL,
system_end TIMESTAMP(12) GENERATED ALWAYS AS ROW
END NOT NULL,
trans_start TIMESTAMP(12) GENERATED ALWAYS AS
TRANSACTION START,
PERIOD SYSTEM_TIME (system_start, system_end),
PRIMARY KEY (emp_name));
```

The only difference to the corresponding syntax in the SQL:2011 specification is the existence of an additional column (in our example `trans_start`), which stores transaction start time. IBM DB2 uses the values stored in this column to track when the transaction first executed a statement that modifies the content of the table. Examples 16 and 17 show the second and third step in defining tables with system time.

Example 16

```
CREATE TABLE v_employees LIKE s_employees;
```

Example 17

```
ALTER TABLE s_employees ADD VERSIONING USE HISTORY
TABLE v_employees;
```

The CREATE TABLE statement in Example 16 creates the new table called `v_employees`, which has the same structure as the `s_employees` table and is used to store old versions of rows. Example 17 modifies the structure of the base table to enable it for versioning and to identify the versioning table.

2) Modifying Data from Tables with System Time

The INSERT statement for tables with system time has the same syntax and semantics as the corresponding statement for system-versioning tables. IBM DB2 supports the same syntax for the UPDATE and DELETE statements as SQL:2011, but the semantics of these operations is different: The modification of a row using the UPDATE statement is maintained so that the new version of the row is placed in the base table and the corresponding old version in the versioning table. Similarly, during deletion, the data from the base table is deleted and copied in the corresponding versioning table. The

system sets the end time of the deleted data in the versioning table to the transaction start time of the DELETE statement.

VI. EVALUATION AND CONCLUSIONS

All temporal data models can be evaluated in relation to several concepts. In this article, we will evaluate the temporal data model introduced in the SQL:2011 specification in relation to the three most important concepts:

- Time dimensions
- Implicit vs. explicit timestamps
- Grouping of time-varying attributes

A. Time Dimensions

The most important concept of temporal data models is time dimension, and there are three different forms of it: valid time, transaction time and bitemporal. Valid time concerns the time when a fact is true in the real world. For this reason, this form of time is independent of its storage in a database and can concern the past, present and future snapshots of the fact.

Transaction time concerns the time the event was present in the database as stored data. The transaction time of an event describes the times, where the event is stored in a database and presents the correct image of the modelled world. Timestamps of transaction time events are defined according to the schedule adopted by the operating system. Therefore, we can build the history of all such timestamps in relation to the past and current time, but not in relation to future. Additionally, only the current values may be updated, and the updates cannot be retroactive.

The union of both forms explained above is called bitemporal. (There is also a special case of bitemporal model, when the valid and transaction times of a fact are identical. As a simple example for this case, the situation where a fact is recorded as soon as it becomes valid in reality can be considered.) The most temporal data models proposed in the literature support only valid time. The specification of temporal data in SQL:2011 supports all three dimensions.

B. Implicit vs. explicit timestamps

The difference between implicit and explicit timestamps concerns how the association of times is represented. In case of explicit timestamps this association is represented by fully explicit timestamp attributes. This issue has consequences in relation to update languages in the following way: While transaction times of facts are supplied by the system itself, update operations in transaction-time models treat the temporal aspect of facts implicitly. On the other hand, the user is responsible to supply valid times of facts. Therefore, updating facts in valid time and bitemporal data models generally must treat time explicitly and are forced to represent a choice as to how the valid times of facts should be specified by the user. The SQL:2011 specification supports explicit and

implicit timestamps in the way described above: In the case of valid time (in application-time period tables), the user is in charge of supplying values for begin and end of the particular time period. The values of transaction time periods in system-versioned tables are automatically set by the system.

C. Grouping of Time-Variant Attributes

Temporal data models support two different approaches in relation how time-variant attributes are attached: tuple time stamping and attribute time stamping. In tuple time stamping, each tuple is augmented by one or two attributes for the recording of timestamps. One additional attribute can be used to record either the time point at which the tuple becomes valid or the time at which the data is valid. Two additional attributes are used to record the start and end time points of the corresponding time interval of validity of the corresponding data. Tuple time stamping is usually applied in temporal relational data models, meaning that the first normal form (1NF) has to be maintained.

The second approach, attribute time stamping, means that the time is associated with every attribute which is time-varying. Therefore, a history is formed for each time-varying attribute within each tuple. As a result, the degree of the relation is reduced by one or two compared with the tuple time stamping, since timestamps are part of the attribute values. Values in a tuple which are not affected by a modification do not have to be repeated. So, the history of values is stored separately for each attribute.

Each of the two approaches has benefits and disadvantages. Tuple time stamping, which implies 1NF may introduce redundancy because attribute values that change at different times are repeated in multiple tuples. On the other hand, temporal relational models can use only this approach. The attribute time stamping overcomes the disadvantage of data redundancy introduced when applying tuple time stamping, but it cannot directly use existing relational storage structures or query evaluation techniques that depend on atomic attribute values. The specification of temporal data in SQL:2011 uses tuple time stamping, because SQL is the language for relational database systems.

D. Deficiencies of the Specification

The main deficiency of the standardized specification for temporal data is the omission to support the PERIOD data type. A period can be defined as a duration that represents a set of contiguous time units within the duration. It has a

beginning and ending bound. The both are defined by the value of two elements: a beginning element and an ending element. Beginning and ending elements can be DATE, TIME, or TIMESTAMP types, but both must be of the same type. The main advantage of the PERIOD data type is that it is naturally (and very easy) to define operations on such a data type. For instance, the following operations are concerned as operations on time periods: CONTAINS, EQUALS, PRECEDES, SUCCEEDS and OVERLAPS.

There are several other deficiencies, which are listed below:

- Coalescing is not supported
- Temporal joins are not supported
- Multiple application-time periods per table are not supported

Coalescing is similar operation to the elimination of duplicates in conventional databases. The aim of coalescing is to merge bring together tuples with identical attribute values and with timestamps, which are adjacent in time, or share some time periods in common. Temporal join means that a row from one table is joined with a row from another table such that their application-time or system-time periods satisfy a condition. (The notion of multiple application-time periods is obvious per se.)

REFERENCES

- [1] Darwen, H.; Date, C.J. - An overview and Analysis of Proposals Based on the TSQL2 Approach, in Date on Database: Writings 2000-2006, C.J. Date, Apress, 2006.
- [2] Kulkarni, K. - Temporal Features in SQL Standard, in <http://metadata-standards.org/Document-library/Documents-by-number>, 2012
- [3] Lorentzos, N. - The Interval-extended Relational Model and Its Applications to Valid-time Databases, in Temporal Databases, 1993
- [4] Melton, J. - SQL:1999, Understanding Relational Language Components, Morgan-Kaufman, 2001
- [5] Saracco, C.M.; Nicola, M.; Gandhi, L. - A matter of time: Temporal data management in DB2, in www.ibm.com/developerworks/data/library/techarticle/dm-1204db2temporaldata/dm-1204db2temporaldata-pdf.pdf, 2012
- [6] Snodgrass, R.T. et al. - TSQL2 Language Specification, in SIGMOD Record 23(1), 1994.
- [7] Snodgrass, R.T. - The TSQL2 Temporal Query Language, Springer Verlag, 1995.
- [8] SQL:2001 Standard - ISO/IEC 9075-2:2011, Information technology - Database languages - SQL - Part 2: Foundation (SQL/Foundation), 2011.
- [9] Zemke, F. - What's New in SQL:2011, SIGMOD Record, 2012.

Analysis of Oral Cancer Prediction using Features Selection with Machine Learning

Fatihah Mohd, Noor Maizura Mohamad Noor

School of Informatics and Applied Mathematics
Universiti Malaysia Terengganu (UMT)
21030 K.Terengganu, Terengganu, Malaysia
mpfatihah@yahoo.com, maizura@umt.edu.my

Zainab Abu Bakar

Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA (UiTM)
40450 Shah Alam, Selangor, Malaysia
zainab@tmsk.uitm.edu.my

Zainul Ahmad Rajion

School of Dental Sciences
Universiti Sains Malaysia (USM),
16150 Kubang Kerian, Kelantan, Malaysia
zainul@kck.edu.my

Abstract—Accuracy is one of the main elements in the disease diagnose. Thus, it is important to select most relevant attributes to generate the optimal accuracy. The objective of this study is to predict more accurately the presence of oral cancer primary stage with reduced number of attributes. Originally, 25 attributes have been identified in order to predict the oral cancer staging. In this study, the integrated diagnostic model with hybrid features selection methods is used to determine the attributes that contribute the most to the diagnosis of oral cancer, which, indirectly, reduces the number of features that are collected from a variety of patient records. Twenty-five attributes have been reduced to 14 features using hybrid feature selection. Subsequently, four classifiers: Updatable Naïve Bayes, Multilayer Perceptron, K-Nearest Neighbors and Support Vector Machine are used to predict the diagnosis of patients with oral cancer. Also, the observations indicate that the Support Vector Machine outperforms other machine learning algorithms after incorporating feature subset selection with SMOTE at preprocessing phases.

Keywords—*diagnose; feature selection; oral cancer; SMOTE;*

I. INTRODUCTION

Early clinical diagnosis is seen as an important element in reducing the mortality rate of deadly disease. The process of clinical diagnosis begins with information gathering or eliciting data from a patient's history. It includes data collection from the patient's primary report of symptoms, past medical history, family history, and social history. In this process, sometimes decision making can be done, where the clinician can start the procedure of formulating a list of possible diagnoses [1]. Then, by doing a physical examination, the physician detects abnormalities by looking at, feeling, and listening to all parts of the body. However, the patient's record is a collection of features and data that leads to problems in the diagnosis.

Another issue is most of the diseases share the same clinical features and scaling. Commonly, a biopsy is taken for the diagnosis. However, the diseases often share many histopathological features as well. Besides that, one disease may show the features of another disease at the beginning stage and may have the characteristic features at the following stages [2,3].

The difficulty to diagnose clinical diseases has attracted many experts to study the solutions from the perspectives of both medical and computer science. A variety of machine learning methods in data mining and artificial intelligence such as feature selection (FS) and classifications are usually applied in the diagnosis of diseases [4,5]. Both FS processes and classification techniques are capable of produce the most

relevant features to build an efficient classifier. In addition, they can also eliminate noise and reduce features to achieve a classification with higher accuracy. Examples of common classification methods include Naïve Bayes (NB) [6,7,8,9], Support Vector Machine (SVM) [10,11,12,13], Genetic Algorithm (GA) [14,15,16], k-Nearest Neighbor (KNN) [17] and Multilayer Perceptron (MLP) [18]. These efficient methods are able to aid doctors in making decision of diagnosis based on the features obtained from the classification. This study aims to produce an efficient predicting diagnosis with deduced number of features that contribute more to the use of oral cancer using feature selection with classification. In this paper, an integrated diagnostic model for selection of the optimum features is proposed. The model is based on integrated a preprocessing phase and hybrid FS which is used to select of features used in the diagnosis process. We also suggested our new hybrid feature selection methods to diagnose the diseases using popular classification techniques such as NB, MLP, SVM and KNN.

Clinical data sets are usually coming with no balance. Class imbalance occurs when one of the classes that are less represented. In the training data, this incident will affect the performance of the algorithm for selecting cases. This often occurs when data collection is not enough [19]. Most classification algorithms aim to minimize the error rate and the percentage of incorrect prediction of class labels [20,21]. To overcome this problem, we propose a preprocessing of imbalanced data set before the features selection stage. In this study, we integrate the Synthetic Minority Oversampling technique (SMOTE) algorithm in our diagnostic model to resolve the problem of imbalance data set.

This paper is organized into four sections. In Section II, the materials and methods included in this study are elaborated. The simulation results of experimental works are presented in Section III. Conclusions are drawn in Section IV.

II. MATERIALS AND METHODS

This section describes oral cancer data set, oversampling method (SMOTE) and features selection algorithms used in this study. The development of the integrated diagnostic model is also presented in this section together with a hybrid feature selection for diagnosis primary stage of oral cancer.

A. Oral Cancer (OC) Data Set

The OC data set in this study consist of 25 variables or features and 82 instances or records [22]. The 25 features are divided into four: (i) demographic features, (ii) clinical signs and symptoms, (iii) histopathological features and (iv) primary stage features (see Table I). The feature of the primary stage is target as class label of disease's diagnostics.

TABLE I. ORAL CANCER DATA SET WITH 25 FEATURES

Demographical Features	Clinical Features
F1: Age	F7: Difficulty in Chewing / Swallowing
F2: Gender	F8: Painless Ulceration > 14 Days
F3: Ethnicity	F9: Neck Lump

F4: Smoking	F10: Loss of Appetite
F5: Quid Chewing	F11: Loss of Weight
F6: Alcohol	F12: Hoarseness of Voice
	F13: Bleeding
	F14: Burning Sensation in the Mouth
	F15: Painful
	F16: Swelling
	F17: Numbness
	F18: Site
	F19: Size
	F20: Lymph Node Involvement
Histopathological Features	
F21: Histological Type / Class	
F22: Differentiation (SCC Type)	
F23: Primary Tumor (T)	
F24: Regional Lymph Nodes (N)	
F25: Distant Metastasis (M)	
Primary Stage (Class/Target)	
One: Stage I	
Two: Stage II	
Three: Stage III	
Four: Stage IV	

B. SMOTE

Clinical data is imbalance in nature, therefore the data need to be preprocessed prior to the next stage of processes. The data set is unbalanced when at least one class have only a small number of instances (called the minority class) while other classes are a majority (with a large number of instances). The limitation of data collection often contributes to imbalance data set [19]. In this situation, classifiers of the majority class usually have good accuracy while the minority class(es) has/have very poor accuracy. In this study, Synthetic Minority Oversampling Technique (SMOTE) algorithm was applied to resolve the problem of imbalance data set during the preprocessing stage. SMOTE is running in a WEKA software environment under the supervised filter function, weka.filters.supervised.instance.SMOTE. The original oral cancer data set must fit entirely in memory. The amount of SMOTE and number of nearest neighbors is specified as Fig. 1.

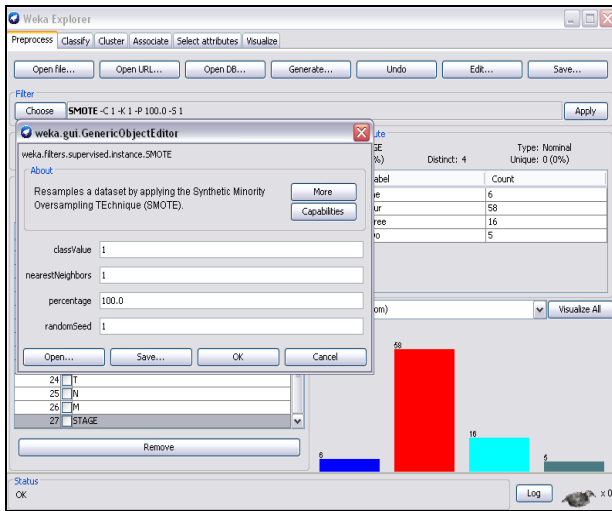


Fig. 1. SMOTE function in Weka software

C. Features Selection

Feature selection (FS) is the process of revealing and reducing unrelated, weakly relevant or redundant features or dimensions in a given data set. The objective of FS is to find the optimal subset. Following are the functions used for feature evaluation (FS) within this study:

- CfsSubsetEval. It evaluates the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them.
- CorrelationVariableEval. It evaluates the worth of features by measuring the correlation (Pearson's) between it and the class. Nominal features are considered as a value by value basis by treating each value as an indicator. An overall correlation for a nominal feature is arrived at via a weighted average.
- InfoGainVariableEval. It evaluates the worth of a feature by measuring the information gain with respect to the class.

All the features were searched using these algorithms:

- BestFirstForward or sequential forward features selection (SFFS). It searches the space of feature subsets by greedy hill climbing augmented with a backtracking facility.
- Ranker. Rank features by their individual evaluations. It is used in conjunction with features evaluators (ReliefF, GainRatio, Entropy, and others).
- LinearForwardSelection with floating forward selection or known as Sequential Backward Selection (SBFS). It is an extension of BestFirst. The search direction can be forward or floating forward selection (with optional backward search steps).

D. An Integrated Diagnostic Model

In this study, the integrated diagnostic model is proposed to diagnose OC data set. It integrates the preprocessing phases and features selection methods (see Fig. 2). The collected OC data are first introduced, as well as the case study with a number of instances and features. The data is preprocessed by scaling or standardizing them to reduce the level of dispersion between the features in the data set. After re-sampling of imbalance data set, the process proceeds to features selection in order to find the most relevant variables in the diagnosis. At this phase, FS techniques are used to select most relevant feature's model, and the various methods of that technique are employed. These models are validated by using the test validation data set. Four algorithms of machine learning are used at this stage to evaluate performance measure accuracy of FS model. Finally, the optimum result gives the best prediction technique or algorithm for that particular type of data set.

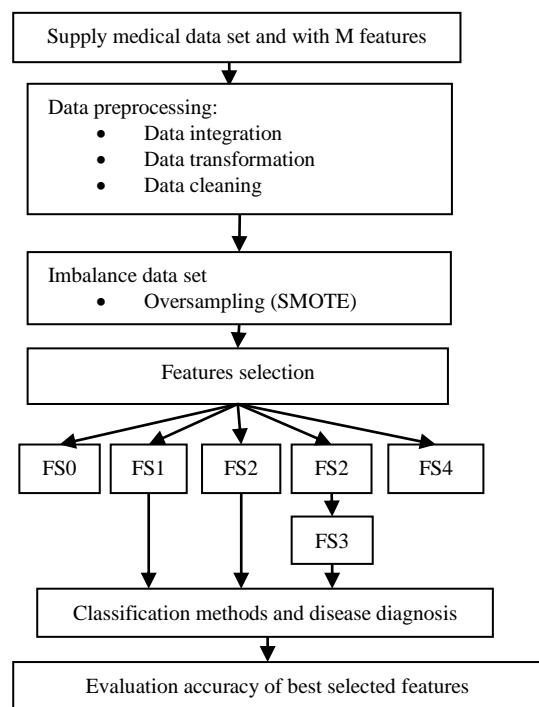


Fig. 2. An integrated diagnostic model for OC data set

III. RESULTS AND DISCUSSION

A. Balance Data set

The original OC data set were categorized into four classes. There were 58 instances of the majority class (stage four), 16 for stage three and stage one and two falls under the category of minority class with the number of instances less than 10. In this study, for the training set 10-fold cross-validation is used. The minority class is over-sampled at 100%, 200%, 300%, and 400% of its original size. Table II shows the result of re-sample an imbalance OC data set using SMOTE. The result after over sampling showed the number of instances is a re-sample to 210 instead of 82 instances.

TABLE II. BALANCED CLASS DISTRIBUTION FOR OC BY APPLYING SMOTE

Class Name	# of Instances	%	# of Instances with SMOTE	%
One	3	3.66	48	22.86
Two	5	6.09	40	19.05
Three	16	19.51	64	30.48
Four	58	70.73	58	27.62
Total	82		210	

Fig. 3 shows the class distribution of each minority class of OC data set, stage one (22.86%) and two (19.05%) are almost balance as majority class, stage three (30.48%) and four (27.62%).

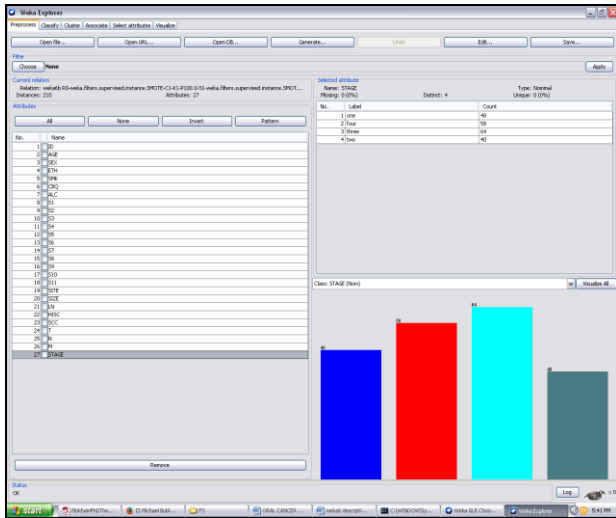


Fig. 3. Balance OC data set using SMOTE in WEKA software.

B. Optimum Features Selected

After loading the data set, the FS algorithms are applied to find the most significant features of the data set. It started with all features selected (FS0), cfsSubSetEval with Best First Forward (FS1), InfoGain Variable Evaluator combined Sequential Backward Selection or known as Linear Forward Selection with Floating Forward Selection (IGSBFS) (FS2), Correlation Variable Evaluator with Ranker (FS3), and hybrid FS3 with CfsSubset Evaluator with Linear Forward Selection (FS4). Table III shows the details of results for each FS method.

TABLE III. SELECTED ATTRIBUTES WITH FEATURES SELECTION METHODS

FS	Method	Selected attributes
FS0	No selected feature	25 attributes
FS1	cfsSubSetEval Best First Forward	2,3,8,9,15,16,17,18,19,20,21,22,23,24 (14 attributes)
FS2	CorrelationAttributeEval Ranker	20,23,21,22,16,19,24,8,2,15,7,17,3,18,5,1,13,9,11,6,25,10,14,4,12 (25 attributes) Remove 11 attributes
FS3	CfsSubsetEval LinearForwardSelection (forward)	20,23,21,22,16,19,24,8,2,15,17,3,18,9 (14 attributes)

FS4	(IGSBFS) InfoGainAttributeEval Ranker	23,21,19,24,20,18,22,16,8,5,1,7,3,2,17,13,5,9,11,12,14,4,6,10,25 Remove gain ratio=0 12,14,4,6,10,25 (6 attributes) Selected features = 23,21,19,24,20,18,22,16,8,5,1,7,3,2,17,13,5,9,11 (19 attributes)
	CfsSubsetEval LinearForwardSelection (floating forward selection)	Optimum features = 23,21,19,24,20,18,22,16,8,15,3,2,17,9 (14 attributes)

The experiment of FS using WEKA software started with 25 features and 210 instances. It ended at FS4 with 14 optimal features namely 2, 3, 8, 9, 15, 16, 17, 18, 19, 20, 21, 22, 23 and 24.

C. Accuracy Classification Performance

The performance measure of accuracy is considered in order to evaluate the efficiency of the FS methods. The measures are compiled by the following unit: Classification Accuracy (%) = (TP+TN) / (TP + FP + FN +TN). In this study, the evaluations are conducted in WEKA with 10 fold cross validation. Four different machine learning algorithms are used to classify the OC data set with four FS methods:

- Updateable Naive Bayes (NB). This is the updateable version of Naïve Bayes and using estimator classes.
- Multilayer Perceptron (MLP). A Classifier that uses backpropagation network to classify instances. This network can be built by hand, created by an algorithm or both. The network can also be monitored and modified during training time.
- SMO-Poly Kernel (E-1.0) (SVM). This implementation globally replaces all missing values and transforms nominal variables into binary ones. It also normalizes all features by default.
- K-Nearest neighbors classifier (lazy.IBk). K-nearest neighbors classifier can select appropriate value of K based on cross-validation. It can also do the distance weighting.

Table IV shows the result for the classifier without oversampling method, SMOTE. It started with select all features of OC data set, 25 features. Next feature selection phase, FS2 is also carrying on with 25 features. Finally, a classifier with 14 selected features from FS3 is generated. Using oversampling (SMOTE), the results for three FS methods with four classifiers show that the features selected by the integrated diagnostic model contributed to improved accuracy of the entire classification algorithm used for the OC data sets.

Table V demonstrates that FS with SMOTE outperforms FS without the implementation of SMOTE. The accuracy of OC data set for FS3 improves from 87.80% to 94.76% for NB,

90.24% to 95.24% for MLP, 86.59% to 96.20% for SVM and 76.83% to 91.43% for KNN. Findings from Table VI are also shown that the highest classification accuracy performance using SVM algorithm, with accuracy of 96.19% with 14 optimal features selection namely 2, 3, 8, 9,15, 16, 17, 18, 19, 20, 21, 22, 23 and 24. The empirical comparison between five FS methods for the entire classifier algorithm is as well performed as graph comparison as Fig 3. It shows the optimal features set from FS3 contribute the highest accuracy performance.

TABLE IV. PERFORMANCE ACCURACY FOR THREE SELECTED FEATURES SELECTION ON OC DATA SET WITHOUT SMOTE

Classification Accuracy Without SMOTE (%)			
Algorithm	FS0	FS2	FS3
NB	85.37	75.61	87.80
	14.63	24.39	12.20
MLP	76.83	79.27	90.24
	23.17	20.73	9.76
SVM	62.20	62.20	86.59
	37.80	37.80	13.41
KNN	75.61	75.61	76.83
	24.39	24.39	23.17

TABLE V. PERFORMANCE ACCURACY FOR THREE SELECTED FEATURES SELECTION ON OC DATA SET WITH SMOTE

Classification Accuracy With SMOTE (%)			
Algorithm	FS0	FS2	FS3
NB	91.90	91.91	94.76
	8.10	8.10	5.24
MLP	94.29	93.81	95.24
	5.71	6.19	4.76
SVM	93.33	93.33	96.20
	6.67	6.67	3.80
KNN	86.19	86.19	91.43
	13.81	13.81	8.57

TABLE VI. PERFORMANCE ACCURACY FOR FIVE FEATURES SELECTIONS ON OC DATA SET

Algorithm	No. of Features	Accuracy (%)			
		NB	MLP	SVM	KNN
FS0	25	91.90	94.23	93.33	86.19
FS1	14	94.76	94.76	92.38	90.95
FS2	25	91.90	93.81	93.33	86.19
FS3	14	94.76	95.24	96.19	91.43
FS4	14	94.76	94.76	92.38	90.95

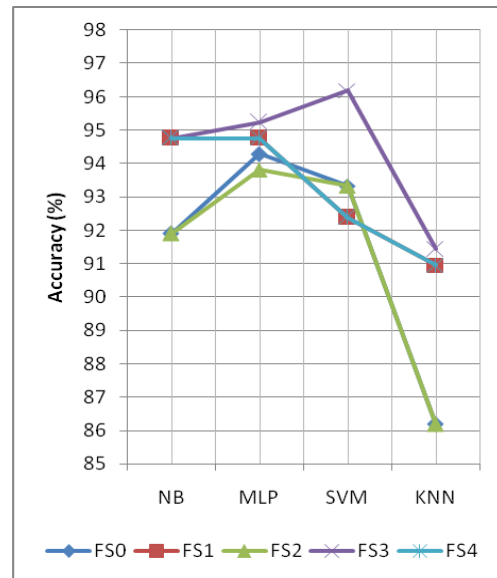


Fig. 4. Performance accuracy comparison between the five features selection methods with NB, MLP, SVM and KNN algorithm.

IV. CONCLUSION

In the field of medical diagnosis, one of the main issues is accuracy in the diagnose of the patient disease. In order to generate the highest accuracy, it is important to reduce and select most related features. Thus, we investigate data reduction methods to be applied in the diagnosis of OC primary stage using machine learning classification methods. In this study, the integrated diagnostic model between preprocessing phases and hybrid FS method to diagnose OC primary stage demonstrated an increase in classification accuracy. It shows highest classification accuracy with 14 optimal features from a set of 25 features. The optimal feature subset was trained with four classification algorithms, Updatable Naïve Bayes, Multilayer Perceptron, K-Nearest Neighbors and Support Vector Machine. Experimental results from this study present that a preprocessing technique before data selection greatly enhances the accuracy of classification. It is also noted that the classifier accuracy enhanced by applied by FS methods than the classifier accuracy without FS. These results clearly demonstrate the great potential of the proposed model for the diagnostic of clinical data.

ACKNOWLEDGMENT

This study has been supported in part of the Exploratory Research Grant Scheme (ERGS) 600_RMI/ERGS 5/3 (3/2011) under the Malaysia Ministry of Higher Education (MOHE) and Universiti Teknologi MARA (UiTM) Malaysia. The authors would like to acknowledge all contributors who have provided their assistance in the completion of the study and anonymous reviewers of this paper. Their useful comments have played a significant role in improving the quality of this work.

REFERENCES

- [1] B. Neville, D. Damm, C. Allen, and J. Bouguot, "Differential diagnosis of oral and maxillofacial disease," in *Oral and Maxillofacial Pathology*, 3rd ed. China: Saunders Elsevier, 2009, Appendix, pp 917.
- [2] J. Xie, J. Lei, W. Xie, X. Gao, Y. Shi, and X. Liu, "Novel hybrid feature selection algorithms for diagnosing erythemato-squamous diseases," in *Health Information Science*, J. He, et al., Eds. Berlin Heidelberg: Springer, 2012, ch. 21, pp. 173-185.
- [3] B. Karlk and G. Harman. "Computer-aided software for early diagnosis of erythemato-squamous diseases," in *Electronics and Nanotechnology (ELNANO), IEEE XXXIII International Scientific Conference*, Kiev, Ukraine, 2013, pp. 276-279.
- [4] L. Li, H. Tang, Z. Wu, J. Gong, M. Gruidl, J. Zou, M. Tockman, and R. A. Clark, "Data mining techniques for cancer detection using serum proteomic profiling," *Artificial Intelligence in Medicine*, vol. 32, pp. 71-83, October 2004.
- [5] K. C. Tan, Q. Yu, C. M. Heng, and T. H. Lee, "Evolutionary computing for knowledge discovery in medical diagnosis," *Artificial Intelligence in Medicine*, vol. 27, pp. 129-154, February 2003.
- [6] I. Rish, "An empirical study of the naive Bayes classifier," in *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, IBM New York, 2001, pp. 41-46.
- [7] S. Mukherjee and N. Sharma, "Intrusion detection using naive bayes classifier with feature reduction," *Procedia Technology*, vol. 4, pp. 119-128, February 2012.
- [8] F. Calle-Alonso, C. J. Pérez, J. P. Arias-Nicolás, and J. Martín, "Computer-aided diagnosis system: a bayesian hybrid classification method," *Computer Methods and Programs in Biomedicine*, vol. 112, pp. 104-113, October 2013.
- [9] M. Wozniak, M. Grana, and E. Corchado, "A survey of multiple classifier systems as hybrid systems," *Information Fusion*, vol. 16, pp. 3-17, March 2014.
- [10] F. Baronti, F. Colla, V. Maggini, A. Micheli, A. Passaro, A. M. Rossi, and A. Starita, "Experimental comparison of machine learning approaches to medical domains: a case study of genotype influence on oral cancer development," in *European Conference on Emergent Aspects in Clinical Data Analysis (EACDA)*, Italy, 2005, pp. 81-86.
- [11] L. H. Lee, C. H. Wan, R. Rajkumar, and D. Isa, "An enhanced Support Vector Machine classification framework by using euclidean distance function for text document categorization," *Applied Intelligence*, vol. 37, pp. 80-99, July 2012.
- [12] G. Orru, W. Pettersson-Yeo, A. F. Marquand, G. Sartori, and A. Mechelli, "Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review," *Neuroscience and Biobehavioral Reviews*, vol. 36, pp. 1140-1152, April 2012.
- [13] B. Zheng, S. W. Yoon, and S. S. Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms," *Expert Systems with Applications*, vol. 41, pp. 1476-1482, 2014.
- [14] D. Mantzaris, G. Anastassopoulos, and A. Adamopoulos, "Genetic algorithm pruning of probabilistic neural networks in medical disease estimation," *Neural Networks*, vol. 24, pp. 831-835, October 2011.
- [15] A. Ozcift and A. Gulen, "Genetic algorithm wrapped bayesian network feature selection applied to differential diagnosis of erythemato-squamous diseases," *Digital Signal Processing: A Review Journal*, vol. 23, pp. 230-237, January 2013.
- [16] S. W. Chang, S. A. Kareem, A. Merican, and R. Zain, "Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods," *BMC Bioinformatics*, vol. 14, pp. 170, May 2013.
- [17] W. L. Tung and C. Quek, "GenSo-FDSS: a neural-fuzzy decision support system for pediatric ALL cancer subtype identification using gene expression data," *Artificial Intelligence in Medicine*, vol. 33, pp. 61-88, January 2005.
- [18] A. E. Hassanien, H. M. Mofteh, A. T. Azar, and M. Shoman, "MRI breast cancer diagnosis hybrid approach using adaptive ant-based segmentation and multilayer perceptron neural networks classifier," *Applied Soft Computing Journal*, vol. 14, pp. 62-71, January 2014.
- [19] J. M. Malof, M. A. Mazurowski, and G. D. Tourassi, "The effect of class imbalance on case selection for case-based classifiers: an empirical study in the context of medical decision support," *Neural Networks*, vol. 25, pp. 141-145, january 2012.
- [20] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Int. Res.*, vol. 16, pp. 321-357, June 2002.
- [21] Q. Wang and W. Chen, "A combined SMOTE and cost-sensitive twin support vector machine for imbalanced classification," *Journal of Computational Information Systems*, vol. 10, pp. 5245-5253, June 2014.
- [22] F. Mohd, Z. A. Bakar, N. M. M. Noor, Z. A. Rajion, and N. Saddki, "A hybrid selection methods based on HCELFs and SVM for the diagnosis of oral cancer staging," in *Advanced Computer and Communication Engineering Technology*, H. A. Sulaiman, et al., Switzerland: Springer, 2015, ch. 77, pp. 821-831.

Decision Support System Utilizing Data Warehouse Technique for the Tourism Sector in Egypt

Tamer A. Abdul-Aziz

Technology Transfer, Director, Technology Innovation Commercialization Office (TICO)
Ain Shams University
Abassia, Cairo, 11566, Egypt
egtato@yaoo.com

Ibrahim F. Moawad

Information Systems Department, Faculty of Computer and Information Science
Ain Shams University
Abassia, Cairo, 11566, Egypt
ibrahim_moawad@cis.asu.edu.eg

Wesal M. Abu-Alam

Tourism Studies Departement, Faculty of Tourism and Hotels
Helwan University
Al-Manial, Cairo, 11792, Egypt
Wesal_alaam@hotmail.com

Abstract- Tourism plays an important role in supporting national economy and creating new jobs. It contributes positively at raising the national income and improving the balance of payments. Egypt depends on the tourism sector to support its national economy. It also represents one of the main sources of hard currency and contributes significantly at solving the problem of unemployment. Notwithstanding, decision makers in Egypt tourism sector have no chance to access a unified data source that can supply information to meet their inquires and expectations. Also the difficulty in conducting analysis and in processing the current data to extract the required information is another challenge because it consumes a lot of time and effort. In this paper, we propose a data warehouse prototype based decision support system for the tourism sector in Egypt. This prototype integrates all the available data sources into a unified data warehouse where data can be viewed, retrieved, and analyzed quickly and efficiently. This system enables the decision makers to access the required information quickly and accurately to support them in making critical decisions at the suitable time.

Keywords- *Decision Making; Decision Support System; Tourism; Data Warehouse; Data Marts; and Galaxy Schema*

1. INTRODUCTION

Tourism is a powerful vehicle for economic growth and job creation all over the world. According to World Economic Forum [1], the Travel & Tourism has continued to be a critical sector for economic development and for sustaining employment in both advanced and developing economies. A strong Travel & Tourism sector contributes at many ways to the development and to the economy of countries. It also makes direct contributions by raising the national income and improving the balance of payments. Authors of [2] clarified that the direct contribution of Travel & Tourism to Gross Domestic Product (GDP) of worldwide in 2012 was 2,056.6 billion USD (2.9% of GDP). This primarily

reflects the economic activity generated by industries such as hotels, travel agents, airlines, and other passenger transportation services. The direct contribution of Travel & Tourism to GDP is expected to have grown by 4.4% to 3,249.2 billion USD (3.1% of GDP) by 2023. Tourism sector is directly and indirectly responsible for generating 261million jobs in 2012 (8.7% of the world's jobs). It is forecasted that by 2023, the Travel & Tourism sector will have supported 338 million jobs (9.9% of total employment), an increase of 2.4% over the next ten years [1].

Egypt depends on the tourism sector to support its national economy. It represents one of the main sources of national income and contributes significantly at solving the problem of unemployment. According to the Egyptian

Ministry of Planning and International Cooperation [3], tourism provides direct jobs for nearly three million people, critical income to more than 70 industries, and 20 percent of the state's foreign currency.

As tourism is considered a composed industry, tourism development plans are associated with many ministries. In Egypt, there is no unified and consistent data sources that can supply information to all decision makers in different authorities for improving the tourism sector. For example, when decision makers in the Tourism Development Authority decide to establish a new hotel in a specific tourist governorate, it is important for decision makers to have information about the number of tourists visiting this governorate, the number of existing hotel units by category, and the number of hotel employees needed and their qualifications. This example shows that the need of acquiring information from various authorities like the Ministry of Tourism, the Egyptian Development Authority, and the Ministry of Higher Education is very important.

Another problem facing the decision makers in the tourism sector in Egypt, as Harb [4] clarified, is exemplified in dealing with large volumes of different valuable tourism data. These data include tourist numbers, tourism nights, percentage of hotel occupations, and the total revenue from the tourism sector at national level, etc. These data are normally stored in hard copies with different formats and in operational databases, which are not easily and timely accessible to decision makers. Harb [4] emphasized that when the president of the Egyptian Tourist Authority (ETA) asks for a report, he has to wait for a substantial time before preparing the report and sometimes he receives data with a very poor quality. Consequently, it is clear that the current situation of the existing data stores in Egypt tourism sector leads to many problems such as:

- Providing inconsistent, inefficient, and poor data.
- Difficulty in collecting, analyzing, and processing the current data to extract the required knowledge and information for decision making.

On view of aforementioned, we proposed a data warehouse prototype that aims at supporting the decision makers in the tourism sector in Egypt by integrating the data sources found in some of the most important ministries and authorities which serve the tourism sector into a unified and a consistent place. We targeted to gather the databases found in (The Ministry of Tourism, the Egyptian Tourist Authority, the Tourism Development Authority, the Ministry of Education, and the Ministry of Higher Education), extract, transform, and load them to a huge data warehouse. Due to data warehouse is a subject oriented, we interviewed some of the officials in the tourism sector in Egypt to determine their needs and requirements.

The rest of this paper is organized as follows. Section 2 reviews the related work. Section 3 shows the adopted methodology to develop the proposed tourism data warehouse prototype. Steps of designing the proposed tourism data warehouse prototype are clarified in section 4, while section 5 explains the steps of building the proposed tourism data warehouse prototype. Section 6 presents a case study. Finally, the paper is concluded in section 7.

2. RELATED WORK

Different approaches were proposed by researchers in the field of supporting decision making in the tourism field. One of these approaches is based on using Decision Support Systems (DSSs). In such approach, some studies [5-7] discussed using DSSs in supporting the tourist to make a decision in choosing the suitable destination matching with his needs and his budget. Other studies focused on using the DSSs for assisting stakeholders in the tourism industry like tourism planners [8], Destination Management Organizations (DMOs) managers [9], tourism demand forecasters [10], and finally tourism marketers [11] to make suitable decisions. Although these studies proposed models related to present solutions to various problems facing stakeholders in the tourism industry, they do not enable stakeholders to instantly analyze and view the required information from different viewpoints and different level of details.

On the other hand, some of these approaches are based on using data warehouse as a tool to support the decision making process in the tourism industry [12-14]. These studies suggested data warehouse models to support decision makers in the tourism industry in various countries like China [12], Romania [13], and Croatia [14]. The nature of the tourism industry in these countries is different than in Egypt (e.g. in the used operational information systems). Moreover, being Egypt one of the developing countries, thus the management information systems are not used efficiently like the situation in a country as China. This leads that the data is scattered and consumes a lot of time to be collected and to be prepared. The research work introduced in [15] is the only study that proposed a data warehouse prototype for the tourism industry in Egypt. Although this work is the pioneer in adopting such approach in the tourism sector in Egypt, the proposed prototype is just a design not a prototype. It is like a guideline on how to build a data warehouse for tourism in Egypt. It is a small data warehouse which relies only on the data existing at the Ministry of Tourism and serves no more than the decision maker at the same ministry. The proposed prototype does not cover the needs of several decision makers in other related sectors like, the tourist employment sector, the tourist learning sector.

To tackle the above-mentioned limitations, we propose a data warehouse prototype for the tourism sector in Egypt covering and including all the aspects of this vital economic field. The proposed tourism data warehouse prototype takes into consideration the nature of tourism industry in Egypt, which depends upon obtaining data from various sources. Moreover, the proposed tourism data warehouse prototype serves different decision makers in the ministries that supply the tourism industry in Egypt. The main purpose of this proposed tourism data warehouse prototype is increasing the efficiency of the decision making process in tourism sector of Egypt through providing a holistic picture of the needed information.

3. TOURISM DATA WAREHOUSE PROTOTYPE DEVELOPMENT METHODOLOGY

Figure 1 represents the proposed tourism data warehouse prototype development methodology that we followed to build the data warehouse. As shown in this figure, the methodology includes three steps:

I. Identifying the prototype requirements and the existing tourism data sources. In this step, we performed in-depth interviews with the tourism sector decision makers. Also, we performed interviews with some officials in certain organizations like the Ministry of Higher Education, the Ministry of Education, and the Center for Documentation of Cultural and Natural Heritage to identify the tourism-related data sources in those organizations, which are very supportive for the decision makers.

II. Designing the proposed tourism data warehouse prototype. In this step, we designed ten subjective data marts, identified the dimensions and measures used in designing these data marts, and designed the galaxy data warehouse schema that builds the proposed tourism data warehouse prototype.

III. Building the proposed tourism data warehouse prototype. In this step, we implemented the galaxy data warehouse schema, executed the ETL process, built the required data cubes, and implemented the presentation layer.

Fig (1) The Tourism Data Warehouse Prototype Development Methodology

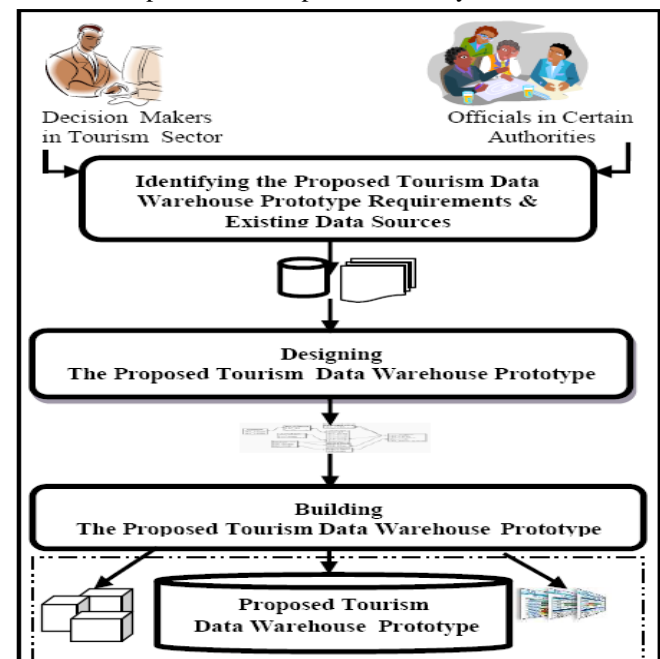
4. DESIGNING THE TOURISM DATA WAREHOUSE PROTOTYPE

To design the proposed tourism data warehouse prototype, the researchers adopted the bottom up design approach, which starts with the prototype requirements elicitation, data marts identification, and finally ends with the data warehouse building. After eliciting the needs and the requirements of the decision makers in the tourism industry, the needed subjective data marts were identified by analyzing those requirements. Accordingly, we identified the dimensions and measures used in designing each data mart. As a result of that, we were capable of designing the galaxy data warehouse schema that builds the complete tourism data warehouse prototype.

According to the needs and requirements of the interviewed decision makers, there are ten subjective data marts. Each data mart holds a specific subject area and performs a specific function such as the indicators of tourists by exported countries, etc. Table (1) demonstrates the ten subjective data marts and their objectives.

TABLE (1) THE SUBJECTIVE TEN DATA MARTS

Data Mart	Objective
1. Indicators of Tourists by Exported Countries.	Counting the number of tourists visiting Egypt in terms of time, region, sub-region, country, mode of transport, and main point of entry.
2. Indicators of Tourists by Visited Cities.	Counting the number of inbound tourists in terms of time, purpose of visit and visited city, governorate, and internal region. .
3. Tourist Nights.	Counting the number of tourist nights in terms of time, exported country, sub-region, and external region.
4. Hotel Indicators.	Counting the number of hotel units, rooms, beds, employees, and occupancy rates in terms of time, city, sub-region, internal region, and existing hotel category and type.
5. Volume of Tourism Receipts.	Measuring the volume of tourism receipts flowed from a particular region, sub- region, and country and on what purpose were spent during a period of time.
6. Tourism Establishments.	Counting the number of tourism establishments and number of its employees in terms of its type and its category according to its place during a period of time.
7. Travel Agencies.	Counting the number of travel agencies and number of its employees in terms of its type according to its place during



	a period of time.
8. Tourist Education Institutions.	Counting the number of tourist education institutions, number of students, number of graduates, and number of staff members in terms of educational type systems and its place during a period of time.
9. Tourist Guides' Indicators.	Counting the number of tourist guides in terms of spoken language and work area during a period of time.
10. Egyptian Heritage Sites.	Counting the number of the Egyptian heritage sites in terms of its type and place during a period of time.

Besides, the domain trees of the tourism data warehouse dimensions show different levels of granularity, and hence the decision makers can easily scroll down and roll up the data marts based on the levels of domain trees. Figure (2) shows the domain trees of Time, City, and Country dimensions.

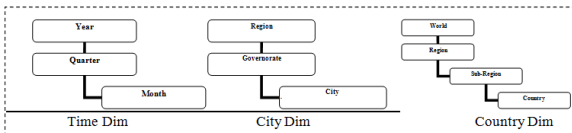


Fig (2) Domain Trees of Time, City, and Country

Finally, designing the galaxy data warehouse schema, which is defined by Poe, et al [16] as a combination of many data marts. The designed galaxy data warehouse schema consists of ten linked data marts to build the proposed tourism data warehouse prototype. Each data mart is designed as a snowflake schema, which consists of a fact table and set of dimension tables. Figure (3) shows a partial view of the designed galaxy data warehouse schema for two data marts: Travel Agencies data mart and Egyptian Heritage Sites data mart. Therefore this partial galaxy schema contains two fact tables (Egyptian Heritage Sites and Travel Agencies) and four surrounded dimensions (Time, City, Travel Agency Category, and Heritage Site Type).

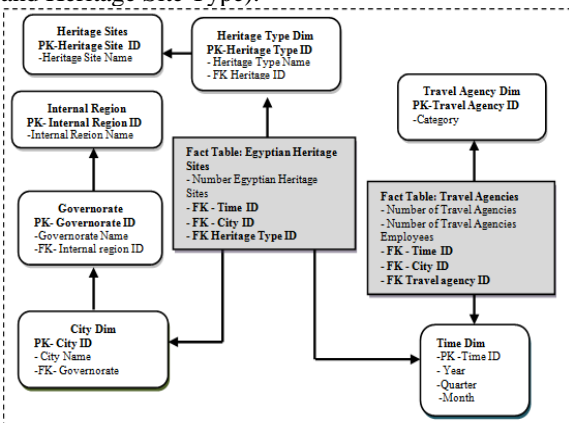


Fig (3) A Partial View of the Designed Galaxy

5. BUILDING THE TOURISM DATA WAREHOUSE PROTOTYPE

To build the designed tourism data warehouse prototype, as shown in figure (4), we firstly performed the ETL process to populate the galaxy data warehouse schema. Secondly, we created the data cubes, which form the ten subjective data marts. Finally, we developed a web based application that contains the data view management, and OLAP database management. This application enables the decision makers to browse the data marts in different multidimensional views, and hence they can generate huge number of reports and charts in a dynamic way.

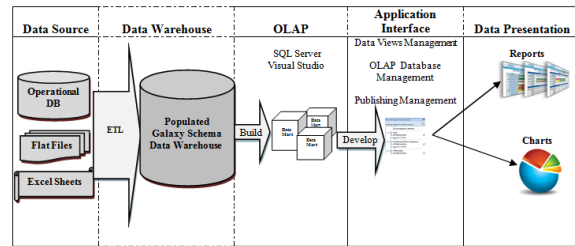


Fig (4) Building the Tourism Data Warehouse Prototype Steps

The proposed tourism galaxy data warehouse schema constitutes of eleven dimensions namely: Time, Main Points of Entry, Country, Purpose of Visit, Hotel Category, Type of Hotel Establishments, City, Tourism Establishments, Egyptian Heritage Sites, Tourist Education and Tourist Guides dimensions. Also, the proposed tourism galaxy data warehouse schema contains ten fact tables namely: Tourists Indicators by Exported Countries, Tourists Indicators by Visited Cities, Tourist Nights, Hotel Indicators, Volume of Tourism Receipts, Tourism Establishments (Restaurants-Cafeteria- Diving Centers), Travel Agencies, Tourist Education Institutions, Tourist Guides Indicators, and Egyptian Heritage Sites.

6. CASE STUDY

To show how the proposed tourism data warehouse prototype is very beneficial for the decision makers in the tourism sector in Egypt, the researchers proposed an illustrative scenario: the decision makers in the Cabinet plan want to set up a Faculty of Tourism and Hotels in a tourist governorate like the South Sinai. This kind of decisions needs many aggregated information like the number of tourist education institutions, the number of students, the number of graduates, the number of hotel units by category, the number of hotel employees, the number of tourism establishments, the number of tourism establishments employees, and the number of travel agencies together with the number of their employees. Having such information can enable decision makers to answer the following questions:

- Is it necessary to build a Faculty of Tourism and Hotels, Tourism Institute or Hospitality Institute in the South Sinai or not?
- If yes, what is the required curriculum?
- How many graduates and staff members are needed?

The proposed tourism data warehouse prototype can provide the decision makers with the needed multidimensional reports (as shown in figures 6 to 9) to assist them to answer the previous questions. Figure (5) shows the number of the existing hotel units by category and the number of hotel employees in the South Sinai governorate in 2011.

Column Labels		
Sinai		
South Sinai		
Row Labels	NO Units	NO Hotel Employees
2011	298	70181
Existing Hotels	298	70181
5 Stars	51	27341
4 Stars	79	23805
3 Stars	68	10214
2 Stars	47	3471
1 Stars	16	679
Under Classification	37	4671
Without Classification	0	0
Grand Total	298	70181

Fig (5) Number of Hotel Units and Employees in the South Sinai by Category in 2011

Figure (6) provides information about the number of tourism establishments and the number of their employees by type in the South Sinai governorate in 2011.

Column Labels		
2011		
Sinai		
South Sinai		
Row Labels	Tourism Establishment NO	Tourism Establishment Employee No
Restaurant	21	200
Cafeteria	92	150
Night Club	4	35
Licensed Diving Centers	116	500
Licensed Bazars	494	800
Grand Total	727	1685

Fig (6) Number of Tourism Establishments by Type and the Number of Employees in the South Sinai 2011

Figure (7) shows the number of travel agencies and the number of their employees by category in the South Sinai governorate in 2011.

Column Labels		
Sinai		
South Sinai		
2011		
Row Labels	Travel Agencies NO	Travel Agencies Employee No
General Tourism	174	568
Ticketing	1	6
Tourist Transportation	30	100
Grand Total	205	674

Fig (7) Number of Travel Agencies by Category and the Number of Employees in the South Sinai 2011

Figure (8) provides information about the number of travel tourist education institutions, the number of students, the

number of staff members and the number of graduates by category in the South Sinai governorate in 2011.

Column Labels				
2011				
Sinai				
Row Labels	Number of Insitutions	Number of Students	Number of Staff Members	Number of Graduates
Schools	4	268	47	64
Higher Institutions	1	355	15	120
2 years	0	0	0	0
4 years	1	355	15	120
Faculties	0	0	0	0
Grand Total	5	623	62	184

Fig (8) Indicators of Tourist Education Institutions in the South Sinai 2011

Based on the previous extracted reports, it is clear that there is no a Faculty of Tourism and Hotels in the South Sinai. There is only a higher institution with number of (355) enrolled students, and (150) graduated students, whereas the number of hotel employees is (70181), the number of tourism establishment employees is (1685), and the number of travel agency employees is (674). This means that the number of graduates does not match the number of available jobs whether in hotels, restaurants, cafeteria, diving centers, or travel agencies. Consequently, there is a shortage in the employment market for this governorate. Therefore, decision makers can answer the previous questions and decide that the Sinai governorate needs a new Faculty of Tourism and Hotels. The decision makers can also determine the required curriculum, students, and staff members.

7. CONCLUSION

In this paper, we proposed a tourism data warehouse prototype for the tourism sector in Egypt. The prototype integrates all the available data sources into a unified data warehouse and provides the decision makers in the tourism sector in Egypt with multidimensional reports that show the required information in different points of view. The case study showed not only the capability of this prototype to issue both integrated and subject oriented reports, but the proposed tourism data warehouse prototype can provide decision makers with reports in various level of granularity as well. The work presented in this paper is an initial step in building a complete data warehouse system with various analytic tools and dashboard indicators to support the decision makers of tourism sector in Egypt.

REFERENCES

1. World Economic Forum. "The Travel & Tourism Competitiveness Report 2013: Reducing Barriers to Economic Growth and Job Creation". Retrieved from http://www3.weforum.org/docs/WEF_TT_Competitiveness_Report_2013.pdf. Author, 2013.

2. World Travel & Tourism Council. "Travel & Tourism Economic Impact 2013". Retrieved from http://www.wttc.org/site_media/uploads/downloads/world_2013_1.pdf. Author, 2013.
3. Ministry of Planning and International Cooperation. "Egypt Development Report- Progressing Towards the Future". Retrieved from <http://www.economist.com/news/business/21577089--turmoil--has--scared>. Author, 2013.
4. Harb, H. Consultant of the Egyptian Tourist Authority's President for Information Technology (Personal Communication, Sep 3, 2008).
5. Asafe, Y., Bnolaji, A., Enaholo, A. and Olubukola, O. "Web-Based Expert Decision Support System for Tourism Destination Management in Nigeria". *International Journal of Advanced Research in Artificial Intelligence*, 2013. PP 59-63.
6. Bunja, D., Bozena, M. and Pavica, N. "Possibilities for Implementation of the Decision Support System in the Croatian Tourism Industry". (Unpublished paper), University of Zadar, Croatia. Retrieved from https://bib.irb.hr/datoteka/344893.Bunja_Krce_Nekic2.pdf. (n.d.).
7. Yu, C. *Personalized and Community Design Support in e Tourism Intermediaries*. Taipei, Taiwan, Springer-Verlag Berlin Heilelberg, 2005.
8. Bousset, J., Skuras, D., Tesitel, J., Marsat, J., Petrou, A., Pantziou, E., Kusova, D. and Bartos, M. "A Decision Support System for Integrated Tourism Development: Rethinking Tourism Policies and Management Strategies". *Tourism Geographies*. 2010, 1 (9), PP. 387 - 404.
9. Baggio, R. and Caporarello, L. "Decision Support systems in a Tourism Destination: Literature Survey and Model Building". *Conference of the Italian Chapter of AIS (Association for Information Systems)*, Verona, Italy, 2005.
10. Petropoulos, C., Patelis, A., and Metaxiotis, K. "A Decision Support System for Tourism Demand Analysis and Forecasting". *Journal of Computer Information Systems*, Athens, Greece. 2003.
11. Wober, K. "Information Supply in Tourism Management by Marketing Decision Support Systems". Retrieved from www.elsevier.com/locate/tourman. 2003, pp. 241 -255.
12. Qioa, X., Zhang, L., Li, N. and Zhu, W. "Constructing a Data Warehouse Based Decision Support Platform for China Tourism Industry". *Information and Communication Technologies in Tourism*, Springer International Publishing Switzerland, 2011.
13. Danubianu, M., Socaciu, T. & Barila, A. "Some Aspects of Data Warehousing In Tourism Industry", *The Annals of The Stefan cel Mare*. Retrieved from www.sciencedirect.com. 2009, PP. 290 - 295.
14. Salmon, B., Marusic, Z. "Data Warehouse as an Organizational tool in Institute for Tourism". *Institute for Tourism*, zagreb, Croatia. Retrieved from http://www.sascommunity.org/seugi/SEUGI1998/marusic_data_warehousing.pdf. (n.d).
15. Hendawi, A., El-Shishny, H. "Data Warehouse Prototype for the Tourism Industry: A Case Study from Egypt". *International Conference on Informatics and Systems*. Cairo University, Faculty of Computers and Information, 2008.
16. Poe, V., Klauer, P. & Brobst, S. *Building a Data Warehouse for Decision Support*. USA. Prentice-Hall PTR. Second Edition, 2008.

Web Crawler System for Distinct Author Identification in Bibliographic Databases

Nancy Dau Marcial Russo Eric Bouwsema Tansel Özyer Reda Alhajj

Department of Computer Science
University of Calgary
Calgary, Alberta, Canada

Department of Computer Engineering
TOBB University of Economics and Technology
Ankara, Turkey

ABSTRACT- A person's name is regularly used to uniquely identify himself/herself from others; unfortunately names are in no way unique and this leads to serious problems. For instance, when trying to retrieve papers from academic database repositories, it can be difficult to distinguish one author from another if the individuals in question have the exact same name. An author can also assume another name, for instance by using the full name. Thus, being able to differentiate which person a specific name is referring to can be tricky. In this paper, we propose a method to solve this ambiguity problem by gathering information from bibliographic databases and using this information to create a social network tree. Based on the relationships created among co-authors it is possible to disambiguate authors with a high-level of accuracy.

Keywords: Namesakes, social network, name ambiguity, academic databases.

1. INTRODUCTION

Names are not unique to a single person. For instance, the most common last name in North America is Smith [1]. Every year there are new lists of top 100 baby names for boys and girls. If one would take a look at these lists, one would notice that the top names do not vary much from year to year.

Name ambiguity occurs at an early age and examples of this can be traced back to kindergarten. If a class has students with the same first name, such as John, the first initial of the student's last name may be included in order for teachers and students alike to differentiate between which John they are referring to: so John Smith would be called John S.

Similarly, if the names extend past just the given name, teachers have to become more creative. The problem with having two people with the exact same name is called a namesake. In order to overcome this issue, the teacher may say that one John Smith would just be called John and the other John Smith would be called Johnny.

This same problem becomes serious when it is transferred to a professional environment, e.g., when it occurs in academic publications that exist in databases repositories like DBLP. Researchers have the problem of namesakes because as we noted before, names are not unique identifiers. The issue with this is that unlike in a classroom where there is a restricted number of students. The ambiguity of names in the academia is a global issue where the same name may exist within the same domain of research or different research interests and within the same university or in different universities, though the latter case is more common.

Unfortunately publications are cited by name and hence it is difficult to identify and separate the exact publications of a given researcher in order to avoid giving the credit where it does not belong.

Digital libraries, e.g., DBLP, IEEE, ACM, Springer, Scopus, etc are all available on the World Wide Web (WWW). It should come to no surprise then that the issue of namesakes and name ambiguity can occur on these sites. Searches can be conducted by looking up authors, but as Figure 1 shows this can lead to confusion, as these websites do not have the ability to filter or identify a single individual from authors with the same name.

In the results of a search for papers written by a common name, say Ken Barker in 2009 alone, we can see the occurrence of namesakes. In Figure 1, we see that DBLP has correctly returned papers written by Ken Barker. The highlighted papers outlined by red and green show that these publications are actually referring to two different persons with the exact same name. When clicking on the bibliographic information on these two authors, we see different information returned. This can be seen in Figure 2.

These academic websites have no way of differentiating between these two authors, so the search results of authors are not accurate. Knowing merely that the author of interest is Ken Barker; these sites will search their database and return any results that have Ken Barker as an author.

While investigating this issue of name ambiguity, we noticed a common trend: researchers throughout their career will meet peers and most of the time will mostly continue to collaborate with them, regardless whether they switch academic institutions. This relationship between authors and

co-authors can help us distinguish one researcher from another.

In this paper, we propose a method that relies on the relationship made between authors of a paper as well as other information taken from bibliographic databases. We have created a web crawler that will go through ACM, DBLP and IEEE libraries. From these sites we collect papers of specific authors as well as information on these researchers.

2009	
124	Derek H. Sleeman, Ken Barker, David Corsar: Report on the Fourth International Conference on Knowledge Capture (K-CAP 2007). <i>AI Magazine (AIM)</i> 30(1):126-127 (2009)
123	Sampson Pun, Amir H. Chinai, Ken Barker: Twins (1): Extending SQL to Support Corporation Privacy Policies in Social Networks. <i>ASONAM 2009</i> :306-311
122	Maryam Majedi, Kambiz Ghazinoor, Amir H. Chinai, Ken Barker: SQL Privacy Model for Social Networks. <i>ASONAM 2009</i> :369-370
121	James Fan, Ken Barker, Bruce W. Porter: Automatic interpretation of loosely encoded input. <i>Artif. Intell. (AI)</i> 173(2):197-220 (2009)
120	Ken Barker, Mina Askari, Mishra Banerjee, Kambiz Ghazinoor, Brennan Mackas, Maryam Majedi, Sampson Pun, Adelepe Williams: A Data Privacy Taxonomy. <i>BNCOD 2009</i> :42-54
119	Adesola Omotayo, Ken Barker, Mostafa A. Hamad, Lisa Higham, Jalal Kawasbi: Answering Multiple-Item Queries in Data Broadcast Systems. <i>BNCOD 2009</i> :120-132
118	Kambiz Ghazinoor, Maryam Majedi, Ken Barker: A Model for Privacy Policy Visualization. <i>COMPASAC 2009</i> :335-340
117	Kambiz Ghazinoor, Maryam Majedi, Ken Barker: A Lattice-Based Privacy Aware Access Control Model. <i>CSE 2009</i> :154-159
116	Sampson Pun, Ken Barker: Privacy FP-Tree. <i>DASFAA Workshops 2009</i> :246-260
115	Shaw Yi Chaw, Ken Barker, Bruce W. Porter, Dan Tecuci, Peter Z. Yeh: A Scalable Problem-Solver for Large Knowledge-Bases. <i>ICTAI 2009</i> :461-468
114	Rashedur M. Rahman, Ken Barker, Reda Alhajj: Performance evaluation of different replica placement algorithms. <i>IJGIC (I)</i> :121-133 (2009)
113	Doo Soon Kim, Ken Barker, Bruce W. Porter: Knowledge integration across multiple texts. <i>K-CAP 2009</i> :49-56
112	Keivan Kianmehr, X. Peng, Chris Luce, Justin Chung, Nam Pham, Walter Chung, Reda Alhajj, Jon G. Rokne, Ken Barker: Mining online shopping patterns and communities. <i>IWAS 2009</i> :400-404
111	Keivan Kianmehr, Shang Gao, Jawad Attari, M. Mushfiqur Rahman, Kofi Akomeah, Reda Alhajj, Jon G. Rokne, Ken Barker: Text summarization techniques: SVM versus neural networks. <i>IWAS 2009</i> :487-491

Figure 1. Namesakes on DBLP when searching for Ken Barker

Text summarization techniques: SVM versus neural networks

Full Text: [PDF](#) [Buy this Article](#)

Authors: Keivan Kianmehr University of Calgary, Calgary, Alberta, Canada
 Shang Gao University of Calgary, Calgary, Alberta, Canada
 Jawad Attari University of Calgary, Calgary, Alberta, Canada
 M. Mushfiqur Rahman University of Calgary, Calgary, Alberta, Canada
 Kofi Akomeah University of Calgary, Calgary, Alberta, Canada
 Reda Alhajj Global University, Beirut, Lebanon and University of Calgary, Calgary, Alberta, Canada
 Jon Rokne University of Calgary, Calgary, Alberta, Canada
 Ken Barker University of Calgary, Calgary, Alberta, Canada

Published in: *IWAS '09 Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services*
 ACM New York, NY, USA ©2009
 table of contents ISBN: 978-1-60558-660-1 doi>10.1145/1806338.1806429

2009 Article
 • Short paper

Bibliometrics
 Downloads (6 Weeks): 12
 Downloads (12 Months): 62
 Citation Count: 0

Knowledge integration across multiple texts

Full Text: [PDF](#) [Buy this Article](#)

Authors: Doo Soon Kim University of Texas at Austin, Austin, TX, USA
 Ken Barker University of Texas at Austin, Austin, TX, USA
 Bruce Porter University of Texas at Austin, Austin, TX, USA

Published in: *K-CAP '09 Proceedings of the fifth international conference on Knowledge capture*
 ACM New York, NY, USA ©2009
 table of contents ISBN: 978-1-60558-658-8 doi>10.1145/1597735.1597745

2009 Article

Bibliometrics
 Downloads (6 Weeks): 5
 Downloads (12 Months): 29
 Citation Count: 1

Figure 2. From the same search results as before we see works from two Ken Barkers, one from the University of Calgary and the other from the University of Texas.

With this information in turn, we can create a network graph/tree. The nodes represent the researchers and within the nodes we have more information related researchers, such as their academic institutions. With these nodes we will

begin to create associations between authors and co-authors. This approach will slowly create a network of clusters among researchers and allow the user to see namesakes or aliases of researchers.

2. RELATED WORKS

As stated in the introduction, we noted that the idea of name ambiguity and namesakes is not a new problem. This topic has been widely investigated and there are several varieties of techniques described in the literature. Looking at the previous research done we see that the idea of duplicating records in large data files was investigated in 1983 [2]. There was research done by Hernandez et. al who gathered large commercial databases and merged data from multiple sources, this he defined as the merge/purge problem and become efficient but costly [3].

Branting has done a comparative study just on name-matching algorithms [4]. Name-matching has been studied and in a study done by Top et al. they showed just how complex name-matching can be with various different situations just based on the name and different alias a person can have, intentional or not [5]. Not only do names differ in spelling, but researchers such as Ji et al. are interested in the way that phonetics can help with name-matching; just another way researchers are thinking outside the box to accomplish the task of matching names to the appropriate persons [6].

Recently, there are researchers who have used multi-layer clustering to try and detect name ambiguity [7]. Jiang et al. used a combination of package-merge algorithm, pattern-matching techniques and fuzzy logic rules in their research.

A study was also carried out by Wu et al. and to solve name ambiguity, they also worked with obtaining more information on the authors, such as workplace and co-author relationships. Wu and his team used this information and applied the association rule and a pre-set threshold to differentiate between name distinctions [8].

Research done in 2005 by Han used the method of K-way spectral clustering, relying on subsequent information given, such as co-authors, paper titles and publication venues [9]. With the clusters, they are able to differentiate groups and decide which groups had which members included in them. Wei et al. on the other hand, concentrated primarily on a biomedical academic website when creating an algorithm for name ambiguity [10]. By using EntrezIDs, he would match the EntrezID information with the information of authors. This allowed them to have a unique ID for each author. Even with this unique identifier and their smaller size database they gained about 75.1% precision when dealing with name ambiguity.

Shin et al. used social networks to resolve the issue of name ambiguity [11]. This research resembles ours described in

this paper, as it focuses on the social networks created by authors and co-authors. They constructed their own namesake and name ambiguity algorithms in order to create their social networks. This research has impacted ours, but it is important to note that they have concentrated on DBLP as their main source of papers; DBLP will connect with other academic sites, but if there is no submission to DBLP then the paper will be left out. Since our research is based on the cumulative returns of the chosen academic research websites, with each having its own crawler, we can gather more information to create our network graph and hopefully also be able to gather more information if one website has a better biographical database.

There are a variety of other techniques used for identification purposes. There are a lot of creative combinations of techniques in just these few sample papers. All of these techniques are used to attempt to solve the problem of personal identification or name ambiguity. We can see that the issue of name ambiguity is a tricky one and all the papers attempted to solve this issue required many steps and more than one algorithm in order to come up with acceptable results.

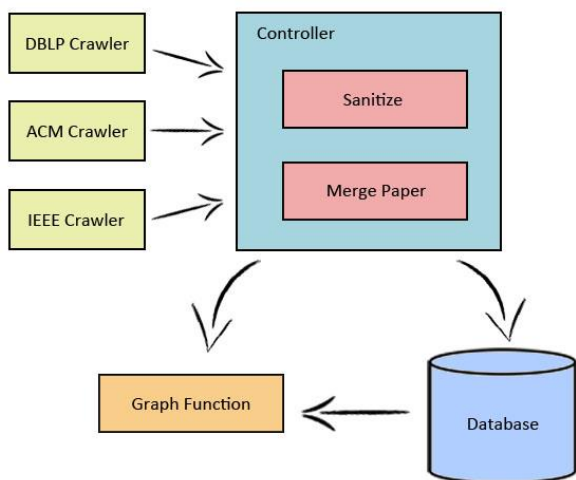


Figure 3. Proposed system architecture.

3. SYSTEM ARCHITECTURE

We need to rely on many parts of the system in order to properly handle name ambiguity. First, it is important to start off with a quick overview of all the parts.

In Figure 3, we see the existence of web crawlers. The current implementation of our web crawlers will comb through the online databases and retrieve papers and authors whose names are attached to these papers. These web crawlers will be scheduled to run or will be manually run by the user. For our current research, we will concentrate on three academic websites: namely DBLP, ACM and IEEE, though others may be easily added if needed.

From the output of the web crawlers we will be able to analyze our results with our controller. It will have two main purposes. The first is to sanitize the information to make sure it is unified and the second is to merge papers that are the same from the different crawlers.

In order to easily retrieve and update the information returned from the web crawler, we will create a database that will store the information.

The graph function that we developed will be our main algorithm to build our graph and the relationships, with which we will be able to deal with namesakes and resolve name ambiguity.

3.1 Web Crawlers

In order to efficiently keep the database up to date, we have created automated web crawlers that will search through our three academic websites. Since all three websites have different bibliographic structures we require different crawlers to return specified information from the underlying database as shown in Table 1.

Table 1. Information returned from web crawlers.

Web Crawler	List of Papers	List of Author(s)	Author Information
DBLP	Yes	Yes	No
ACM	No	No	Yes
IEEE	No	No	Yes

Since DBLP contains all of the author’s papers, it will be our main source to get all authors and all the paper’s information. Unfortunately DBLP doesn’t have any of the author information such as Affiliation or email. This information is useful to identify authors with same names, but in this paper we will concentrate mainly on comparing an Author’s co-authors to distinguish them.

With the list of papers and authors returned, the system will be able to utilize the information to help us create a network graph.

3.2 Controller

In the second step of our system architecture, we have the controller. It has two main functions: to merge papers returned from the crawlers that are the same. The crawlers will be pulling from DBLP and information from ACM/IEEE; there is possibility that researcher groups could have submitted their papers to individual academic websites. The paper could have been approved for more than one submission. If there are multiples of the same paper the controller will identify this and merge them together along with the information. The second role of the controller is to be able to sanitize the information. What is meant by sanitize, is to clean up the results so that our algorithm will

be able to effectively analyze the information without any issues, such as issues which may arise when the results include special characters, these can be seen in languages that use accents such as French or Spanish.

3.3 Data Store

In the database we will be storing two sets of information taken from our controller.

The first will be the names of the authors. We will keep a list of names, which will be taken from the results of the DBLP crawler. The names will be added one by one into the database. In the beginning, we will have only one author; from there we will search the co-authors of a given paper. Through this iterating process, we will eventually be searching those previously added authors to see if there is a name that currently does not exist in our list. We can see an example of how this will be stored in Table 2.

Table 2. Table kept by the database with the list of authors currently seen.

Name
Ken Barker
Reda Alhajj
Bruce Porter

This will be used by both the controllers and by the graph function in order to create relationships. When we come across a new author we will add him or her to the current table of names and then in a next iteration of the author relationship table we will look for new relationships amongst them. This will create a domino effect and spread through the full database, returning to us a complete result. An example of how this is done is given in Table 3.

Table 3. Relationships between authors and co-authors

Relationship
Ken Barker – Reda Alhajj
Ken Barker – Bruce Porter

Looking at the tables, we can see how this relationship among co-authors will help us building our network graph. We see from the tables that there is a relationship between Ken Barker and Reda Alhajj, and another relationship exists between Ken Barker and Bruce Porter. Also note, that there is currently no relationship between Reda Alhajj and Bruce Porter in our database. This lack of a relationship is just as important as the relationships that exist because they will help us create an accurate network graph moving forward. Indeed the lack of relationship may be a good indicator that the two occurrences of Ken Barker are not the same person.

The second information that the database will keep track of is the list of paper titles. Once again, the system will be merging those papers that are exact duplicates of one another. Unlike other research papers, this one is based on multiple academic databases so the occurrence of duplicate papers is likely since these are separate websites. We will have to be able to properly identify and merge papers that are similar.

3.4 Graph Function

Using our knowledge of the current system, we now move on to the graph function. This part of the system is to be considered the heart of our overall architecture. The role of the graph function is to take the knowledge from the database and turn it into a network graph of clusters. From these clusters we will be able to differentiate namesakes as well as be able to eventually merge alias names.

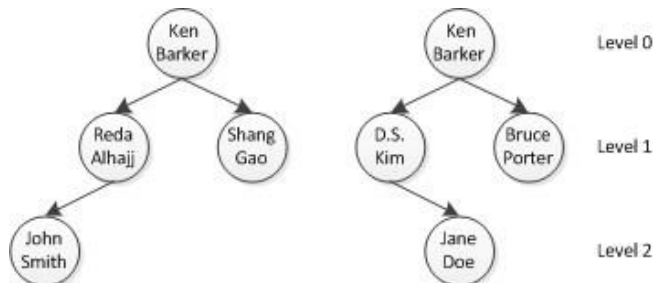


Figure 4. Visualization of the network graph

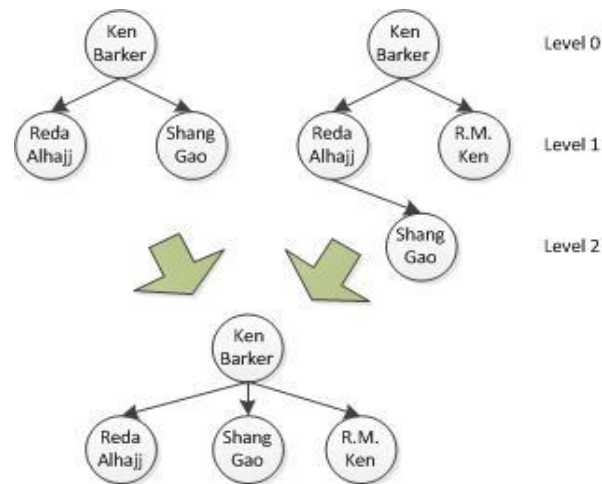


Figure 5. Merging of two clusters in the case of Namesake.

3.4.1 Namesake

We can see from Figure 4 how this graph function will be developed. A node will represent the authors and the relationships are based on co-authors. So, if a node or author has collaborated with another node, we call this a relationship and connect those two nodes together. This will basically become a forest of trees. In Figure 4, we see that

two Ken Barkers exist. They both have the exact same name, but their relationships are distinct from each other. This will be the easiest case to confirm that these authors are indeed separate authors and that there is no need to merge them.

If for example in Figure 5 there is a case where two Ken Barkers exist with the same name and the co-authors are relatively similar, we will have to look at what degree of similarity these two clusters have. This will depend on the co-authors and the levels that these similarities occur on. If the similarities threshold is met, we will combine the two clusters.

3.4.2 Six Degrees of Separation

Another case study that needs to be explored is the idea of having a namesake within a namesake. In our tree, we note levels on the right. The deeper we move downward into the tree, the less value we give to that relationship. This rule is to prevent a phenomenon known as the Six Degrees of Separation [12] (or in pop-culture, the Seven Degrees of Kevin Bacon). This idea is well known: if one picks a person in the world, usually a celebrity, it will take six people or less in order to “know” this person. This will create a lot of “friend of a friend” instances, but in the end one will somehow be socially related to that chosen person.

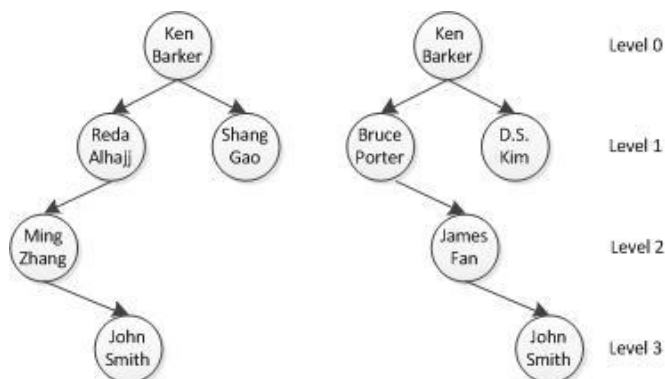


Figure 6. Theory of degrees of separation applied.

Unlike the idea of Six degrees of separation, we do not have a database that includes all of society and because of this we must put a limit on how much of a degree of separation we can allow before our analysis will no longer be optimal. In Figure 6, we added another level of our network graph. Note that on level 3 we have two nodes in separate clusters with the name John Smith. Looking at the two clusters, we see there is noticeable difference in the other co-authors that Ken Barker has worked with. So, even though Ken Barker has associations with John Smith, this will have little effect since it occurs lower on the tree and the co-authors before him do not meet the threshold of similarities. We want to restrain our tree from adapting to the idea of six degrees of

separation, so we have reduced the number into half; we will look at 3 degrees and put less weight on each level.

3.4.3 Multiple Names (Alias)

Another example to explore is the idea of an author using more than one name when publishing papers. Examples of this can be seen when a person uses his or her middle name to avoid the issue of namesakes. Or if the individual moves to another institution and the spelling of the name uses special characters. When first iterating through the tree, we may have two separate nodes for one author.

In the case of two nodes that actually represent one author occurring as shown Figure 7, the system will have to be able to first identify and then merge the clusters. If the two clusters have a high enough agreeability rate, the root of the smaller cluster will have to be sanitized to conform to the larger cluster, which we will be merging with. This sanitization process is the only difference between a regular merge versus this specific type of merge. The clusters will be combined by integrating them together. In order to keep the databases as similar as possible, we will also have to reassign the papers that were previously assigned under the old author’s name to the new author’s name. This creates a cycle between the controller, the database and the graph algorithm. We can see the example of a merge occurring presented visually in Figure 7.

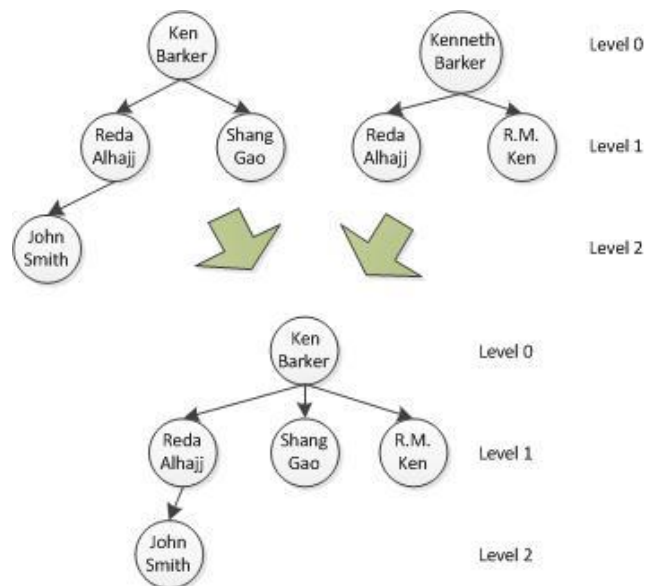


Figure 7. Merging of two separate clusters when authors use alias.

With the current system this situation is not handled. The issue with our current implementation is that if an author were to move to another academic institution, and no longer write with their former peers, the system then has no way of connecting the two entities. A solution to this issue is to be able to go through the paper using a PDF reader.

Unfortunately, the implementation for the PDF reader and the current system has not yet been incorporated in this running system. This is discussed more in future work.

4. EXPERIMENTAL RESULTS

4.1 Setup

The results are highly dependent on the system’s ability to obtain all of the available papers from the specified author, using the three implemented web crawlers. The system must also have the ability to distinguish different authors with the same name. DBLP has standardized their data, thus obtaining authors and their papers do not require a check to be made for name abbreviation, or modifications; hence all records can be obtained by using our DBLP crawler to search the database.

The first step in the system is to declare a set of authors to act as our seeds. The crawlers will then be used by using the previously declared set to perform a DBLP search. Since DBLP returns a list of papers with their authors, we break them into two objects. The paper: this object consists of the paper’s title, year, authors, topic and where it can be found, the hyperlink; this is shown in Figure 8.

The second object that is created is the author: this consists of the authors name, affiliations, email, coauthors, and papers that he or she has participated in writing; this is shown in Figure 8.

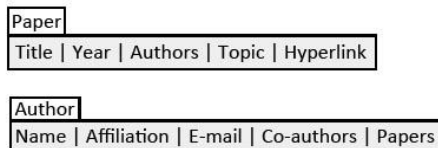


Figure 8. Representation of the two objects created.

Once all the available information of an author has been collected, the program enters a cyclic stage, the co-authors now become seeds and the cycle starts again. Eventually when all seeds have being added to the main set, we have finished collecting information and can now move to building our network.

The program then moves on to the next stage: Identifying authors with the same name and merging them into one, as shown in Figure 5. If the situation occurs that we have a namesake within a namesake, the system will have to perform the most complicated step of the identification algorithm by using up to three levels of our created network to correctly merge authors and networks, as shown in Figure 6. If two existing networks are separated and no affiliation is made between them up until three levels, we do not merge. If there does exist a co-author within three levels we will assume this is the same person and merge.

Table 4. Results for seed Reda Alhajj

	Distinct Papers	Distinct Co-Authors
Initial Results	291	698
Merge based on Names & Papers	24	197
Path Merge	13	197

4.2 Results & Analysis

Our system was able to successfully reduce the number of potential distinct authors for two test names (Reda Alhajj and Ken Barker) drastically. In the case of Reda Alhajj we have a single author who is prolific and has numerous co-authors.

From our initial results returned from DBLP we were able to group first into 24 distinct authors, and then through path operations we were able to reduce down to 13 distinct authors. This drastically reduces the number of potential candidates to match using additional heuristics. (Table 4.)

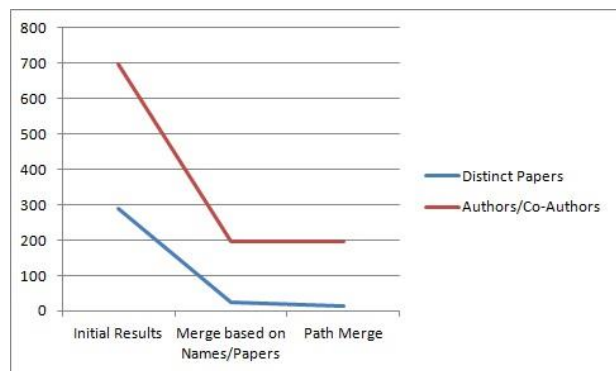


Figure 8. Results for seed Reda Alhajj

Table 5. Results for seed Ken Barker

	Distinct Papers	Distinct Co-Authors
Initial Results	138	137
Merge based on Names & Papers	23	135
Path Merge	22	135

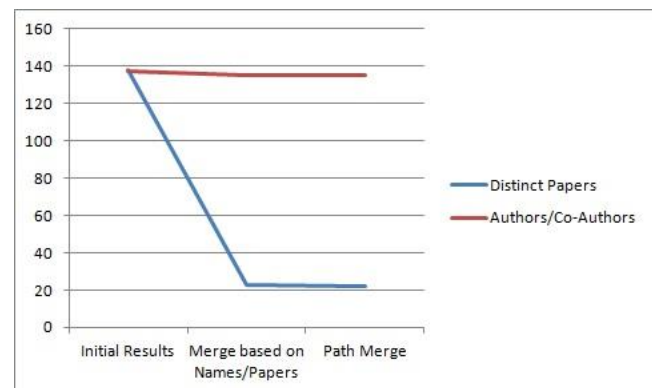


Figure 9. Results for seed Ken Barker

Our second author, Ken Barker, consisted of two distinct authors who have similar names. Our initial results returned 138 potential authors and we were able to reduce it to 23 potential candidates. The path procedure did not provide drastically improved results, only reducing the number of potential candidates to 22. (See Table 5.)

When examining the results for Ken Barker we can see that there are a number of distinct clusters who consist of most of the collaborative work done by both individual authors. For Ken Barker currently at the University of Calgary you can see a number of past and current students and staff who are the co-authors. For Ken Barker from University of Texas this also holds.

Table 6. Clusters for authors ‘Ken Barker’

Ken Barker Univ. of Calgary/Univ. of Manitoba/Univ. of Alberta	Co-Authors	Papers
Brenan Mackas, Jawad Attari, Philip W. L. Fong, Kofi Akomeah, Angela Cristina Duta, Walter Chung, M. Sheelagh T. Carpendale, Justin Chung, Nelson C. N. Chu, Chenen Liang, M. Mushfiqur Rahman, Rosa Karimi Adl, George Shi, X. Peng, Adepele Williams, Christoph W. Sensen, Chunyan Wang, Janaki Gopalan, Jamal Jida, Leanne Wu, Nancy Situ, Maryam Majedi, Kambiz Ghazino, Reda Alhaji, ...	54	79
Moustafa A. Hammad, Jalal Kawash, Adesola Omotayo, Lisa Higham	4	4
Joseph Osuji, Faith-Michael E. Uzoka, Okure U. Obot	3	1
Dina Said, Peter Federolf, Lisa Stirling	3	1
Randal J. Peters, Coimbatore Rajagopal Saravanan	2	2
Peter C. J. Graham	1	1
Ahmad R. Hadaegh	1	2
John Aycock	1	1
C. I. Ezeife	1	2
Wendy Osborn	1	2
M. Tamer Özsü	1	2
Ramon Lawrence	1	3
Subhrajyoti Bhar	1	2
Amin Y. Noaman	1	1
Sergio Camorlinga	1	2
Sylvanus A. Ehikioya	1	3
Md. Moniruzzaman	1	1
Michael Zapp	1	1
Ken Barker Univ. of Texas/Univ. of Ottawa	Co-Authors	Papers
Pedro Romero, Mark Greaves, Daniel Hansch, Rutu Mulkar-Mehta, Michael Eriksen, Andrés Rodríguez, David Gunning, Bhalchandra Agashe, Blake Shepard, Michael Glass, Moritz Weiten, David D. McDonald, Nancy Salay, Gavin Matthews, Jing Tien, Bonnie E. John, Benjamin N. Groszof, Paul G. Allen, Eduard H. Hovy, Sourabh Patwardhan, Jérôme Thoméré, Doo Soon Kim...	50	21
Sylvain Delisle, Terry Copeck, Stan Szpakowicz	3	4
David Corsar, Derek H. Sleeman	2	2
Nadia Cornacchia	1	1

Interestingly enough for both Ken Barkers the system does not care that the author has moved universities, but rather groups based on collaborative efforts, where colleagues continue to work together even after moving to new institutions. (See Table 6.)

An issue we have noticed with this system though is when two authors write a paper, but one or the other of the authors does not collaborate with any other authors at the time of crawling the database. As you can see with Ken Barker (Univ. of Texas) and Nadia Comacchia, with only one paper and only one co-author (Nadia Comacchia only has one paper on DBLP) it is hard for us to cluster this author, and thus requires us to use other methods to determine which Ken Barker this is. For the purpose of Table 6 we were able to determine the proper author based on CV’s and not using our system.

While we are not able to determine with 100% accuracy the clustering of each author, we have shown that we can drastically reduce the number of potential unique authors using our system. Building upon this work we should be able to determine with a high level of accuracy each distinct author.

5. FUTURE WORK

A current issue that can be resolved with future work is that bibliographic information is not as consistent as it is needed to be among the websites. Since websites such as DBLP, IEEE and ACM may run independently of each other, the bibliographic information provided range from good bibliographic information, to basically no bibliographic information; this makes it more difficult to pull from websites as they are not all consistent with each other.

For future works what can be done with the current system to gain better accuracy would be to use a PDF reader to gain access to information such as email or the educational institution. This will allow future research to more accurately connect the authors with this information and thus return stronger results than what we’ve been able to do thus far. With ACM and other academic research websites currently enforcing stronger layouts of their submitted papers this means that such problems as namesakes can be better dealt with. For example, just as this paper, many other papers are forced to now include author information such as e-mail and the institution which the researchers are writing for. E-mail is a unique identifier as no two people can ever have the same e-mail under the same domain name. Furthermore a researcher is not allowed to be working for two or more academic locations at one time, from this we can safely say that this is also a unique identifier, as published works of an author can only be from that single institution. This would help us be able to deal with authors who may write under several different names. Also for better

accuracy to distinguish authors a restriction to comparing common names could be placed. For example if both authors share a paper with “John Smith” since this is such a common name it should not be used to identify the uniqueness of the author.

The current system also runs a lot slower than originally anticipated. A suggestion for future work is to incorporate something such as Hadoop [13]. This open-source framework will allow the system to be reliable and scalable. Since it is utilizing distributed systems this will help speed up the system and will be able to get faster results.

6. CONCLUSION

In conclusion, we proposed a solution to solve the problem of name ambiguity. Through research we suggested using a method that relied on networking. Our data was taken from web crawlers that searched through academic websites, and extracting information specifically from the bibliographic pages that they supplied.

We considered such things as namesakes and proposed a theory for how to deal with authors who may use multiple names when publishing papers. Our main goal was to be able to distinguish papers by researchers who publish at the same time as someone else who has the same name. Our proposed system is able to make a network from co-authors providing us associations that we can use to properly distinguish these authors who have the issue of namesakes.

We have also provided future work that can be done, in order to improve our current system, this can be found in the future works section of our paper.

REFERENCES

- [1] Wikipedia List of most common surnames in North America.
http://en.wikipedia.org/wiki/List_of_most_common_surnames_in_North_America.
- [2] Bitton D. and DeWitt, D., Duplicate Record Elimination in Large Data Files, ACM Transactions on Database Systems, pp. 255-265, 1983.
- [3] Hernandez, M. and Stolfo, S. The merge/purge problem for large databases. In Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 127-138, 1995.
- [4] Branting, L.K., A comparative evaluation of name-matching algorithms. ICAIL '03 Proceedings of the 9th international conference on Artificial intelligence and law.
- [5] Top, P., Dowla, F. and Gansemer, J. A Dynamic Programming Algorithm for Name Matching. Computational Intelligence and Data Mining, 2007. CIDM 2007. Page(s): 547-551.
- [6] Ji, H., Grishman, R. and Wang, W. Phonetic name matching for cross-lingual Spoken Sentence Retrieval. Spoken Language Technology Workshop, 2008. SLT 2008. IEEE. Digital Object Identifier: 10.1109/SLT.2008.4777895. Publication Year: 2008, Page(s): 281-284.
- [7] Jiang, W., Wang, A, Wu, C., Chen, J. and Yan J. Approach for Name Ambiguity Problem Using a Multiple-Layer Clustering. Computational Science and Engineering, 2009. CSE '09. Volume: 4, Page(s): 874-878.
- [8] Wu, B., Cai, W. and Li, Y. Association analysis and case study framework based on the name distinction. Computer Application and System Modeling, 2010. Volume: 4, Pages V4-285 – V4-289.
- [9] Han, H., Zha, H., and Giles, C.L. Name disambiguation in author citations using a K-way spectral clustering method. Digital Libraries. JDCL '05. Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Object Identifier. pp.334-343, 2005.
- [10] Wei, C., Huang, I., Hsu, Y. and Kao, H. Normalizing Biomedical Name Entities by Similarity Based Inference Network and De-ambiguity Mining. Bioinformatics and Bioengineering. BIBE '09. Ninth IEEE International Conference on Digital Object Identifier. pp.461-466, 2009
- [11] Shin, D., Kim, T., Jung, H. and Choi, J. Automatic Method for Author Name Disambiguation Using Social Networks. Advanced Information Networking and Applications (AINA), 2010 24th IEEE International Conference on Digital Object Identifier. Page(s): 1263-1270.
- [12] The Guardian: Proof! Just six degrees of separation between us.
<http://www.guardian.co.uk/technology/2008/aug/03/internet.email>.
- [13] Welcome to Apache Hadoop! <http://hadoop.apache.org/>.
- [14] Ferreira A. A., Gonçalves M. A. and Laender A. H. F., A Brief Survey of Automatic Methods for Author Name Disambiguation, SIGMOD Record, Vol. 41, No. 2, June 2012.

Technical object as a system of “growing” structure

Marek Młynczak

Wroclaw University of Technology
Faculty of Mechanical Engineering
Chair of Logistic System Operation,
Transportation System Operation and Hydraulic Units
ul. Wyspianskiego 27
50-370 Wrocław, Poland

Abstract: Modern technical objects are complex systems of multilevel hierarchical structure. In the literature, problem of system structure, its catalogues and spare parts catalogues, as well as formal definition of the structure is rarely undertaken. Failure analysis of technical objects shows that at reliability field test, the knowledge about its structure is not always a priori required. It leads to the concept of “growing structure” which is created continuously while operational events appear. Formal definition of hierarchical structure compared to growing one is in the paper presented.

Keywords—technical object, system, reliability field test, construction structure

I. INTRODUCTION

The complexity of modern technical objects in operation enforces the needs of creative use of operation performance theory, especially in the field of study of these processes for operation rationalization of these objects. An important issue of process management operation is operational data collection creating the base of knowledge about the system and its operation. Information base (knowledge base) is developed in the process of operation using the diagnostic methods, data verification and processing up to final assessment of machine effectiveness. The database is created in order to use it in decision making process and should contain information about operational states, time, failures identification, etc. Machine system operators should collect all these data in the simplest and reliable way but also protecting data from losing it.

During the operation of the facility, many physical transformation processes of energy and mass take place. It is accompanied by the aging and degrading processes leading to a total or partial loss of its performance. This justifies the need of performing diagnostic tests in the operation and collection of data documenting the variability of these properties over time. Systems for collecting and processing information about events are essential for optimal management of a process operation of the facility.

In machine operation an information base used to control the operation process of the object forms a set of data (database, knowledge base) with the algorithms of processing, transmission and storage of the data.

Also in the earlier, design phase, technical objects should satisfy several technical, economical and quality requirements. Some of the most important are those, which have an effect on: availability, safety and operational costs. Most of computer aided design systems help in creating technical object, what is based on data covering attributes ensuring: consistency,

dimensional synchronization and functionality. The purpose of this paper is to analyze the possibility of applying operational, reliability and safety data into database in more effective way. Random events disturb proper process and lower effectiveness of operation through undesired failures, hazardous events and other economic consequences. Object to be designed is modelled in systemic approach that considers elements described by attributes due to functionality, reliability and safety.

An important element of the information system operation management is an algorithm of dynamic information processing of its functional structure based only on observed events in real operation. This algorithm is an ordered set of instructions of input data transforming into information to make it useful in recognizing an item or object module (source of an event) in its functional structure. Data processing algorithm in the collection of information for decision-making is described as mathematical decision-making model.

The presented information encoding algorithm about events in the operation of a technical object creates a knowledge base about its structure and is an important element supporting the process of its operation.

Created knowledge base allows for rational use of the potential capacity of the facility and its individual components, by reducing the risk of errors resulting from the lack of the necessary information on how to use the principles handling, material supply and economically reasonable management [13].

II. MODELING OF COMPLEX TECHNICAL OBJECTS

The problem of modelling complex technical objects appeared in reliability theory in the early publications of Birnbaum, Saunders, Barlow and Proshan, where the object is defined as a system composed of elements [2]. The system structure is most often treated as a hierarchical structure,

assigning sets of components to a higher level of decomposition by certain relations [3, 4, 12, 15].

A review of the many definitions of the system shows that “elements ordering is the weakest point, structure definition requires establishing the existence of system elements to be in certain order, relationships and linkages between structure and function of the system” [15]. Next, the author proposes a definition of the structure as a “generalized characteristics of the specific properties of the system, fixing in the abstract form elements, relationships, feedbacks, their arrangement and organization.”

Object structure is the basis for modelling an operational database in the research of real mechanical object [9]. Quoted here are some examples of hierarchical, multi-level structure of the object and relevant documents to collect data on failures and other operational events.

Most studies of complex technical systems are based on the two-level decomposition where the system is made up of elements attributed with a hierarchical system of codes [9, 13, 16, 17].

In the systems with redundancy it is taken into account the complex structure of the object by assigning redundant elements two-level code. Sample block A of higher level is decomposed into elements A1, A2, A3 having code of higher level A and $i=1,2,3$ of lower one [11].

Creating structure of technical objects is done usually using a hierarchical approach; however, it is mostly intuitive process. Catalogues of spare parts are built by assigning elements to components, further to the subassemblies, assemblies, units and systems, which is a typical example of a hierarchical structure [10]. An example of v-belt stretcher subassembly of cooling subsystem of diesel engine is shown in Fig. 1. Elements are numbered 1 to 25 and correspond to higher decomposition level (cooling subsystem).

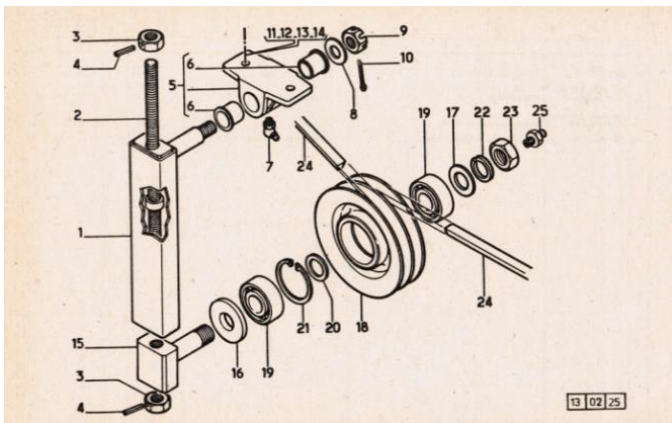


Fig. 1. An example of v-belt stretcher subassembly – catalogue page

III. THE ALGORITHM OF SYSTEM DECOMPOSING

Technical system is characterized by a hierarchical structure consisting of set of elements and its properties and is described in the form of triplet (1) [5, 7, 8]:

$$OM = \langle E, W, R \rangle \tag{1}$$

where: $E = \{e_i\}$, $i = 1,2, \dots, n$ - a set of elements (subsystems, components),

$W = \{w_{is}\}$, $s = 1,2, \dots, m_i$ - a set of distinguished properties (attributes) of elements,

$R = \{R_l(w_{is})\}$, $l = 1,2, \dots, r_i$ - a set of relations assigned to a set of properties of elements.

As time passes, elements of the object appears in “information process” giving in time section sign (event), that operator should react on it. Information regards one of many attributes attached to the element (Fig. 2).

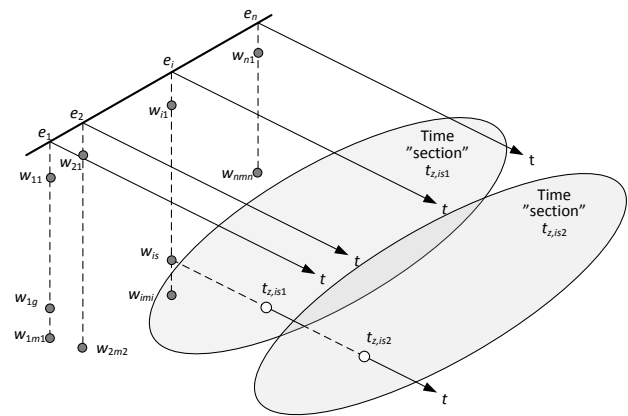


Fig. 2. Time sections with regard to events created by i -th element

The system is subjected to a multi-level decomposition into subsystems, in practice, called systems, units, assemblies, subassemblies, components, and so on. Later, the name of element is reserved exclusively for single-element module, referred to as non-disintegrable. The first level of decomposition is the main division, and it forms n_1 main modules (2):

$$E_{j_1}; j_1 = 1, 2, \dots, n_1 \tag{2}$$

Since each of the components E_j has to belong to only one main module and so their set sum gives the set of all elements (3), i.e.:

$$\bigcup_{j_1=1}^{n_1} E_{j_1} = E \tag{3}$$

If for some $j_1 = 1, 2, \dots, n_1$ module E_{j_1} is composed of only one element e_i , i.e. $E_{j_1} = \{e_i\}$, then the module is no longer decomposed and element e_i is identified as the only element of the module E_{j_1} . Especially, when the number of units is equal to the number of elements, namely $n_1 = n$, then the system is completely decomposed at the first level, and it ends up decomposition process. If $n_1 < n$, then decomposition process continues to decompose the second level. Let n_1^* denotes the number of single-element modules decomposed at the first

level of the system, i.e. non-disintegrable. These items are excluded from the further process of decomposition. Then, the number $d_1 = n_1 - n_1^*$ of multi-element modules remains to be decomposed further at lower level. Decomposition of each multi-element module proceeds in the same way as at first level and is continued until all multi-element modules are decomposed.

Due to the finite number of elements of the system, decomposition process finally completes and the number of levels of the system depends on the number of its components and the complexity of the functional structure. Let's introduce further designations:

l – number of decomposition levels ($l \in \mathbb{N}$), \mathbb{N} is the set of natural numbers,

n_i – number of modules on the i -th decomposition level of the system ($1 \leq i \leq l, n_i \in \mathbb{N}$),

n_i^* – number of single-element modules on the i -th decomposition level,

d_i – number of multiple modules on the i -th decomposition level.

Of course, the equality (4) exists:

$$\sum_{i=1}^l n_i^* = n \quad (4)$$

The total number of decomposition $d = \sum_{i=1}^l d_i$ is (5):

$$d = \sum_{i=1}^l n_i - n \quad (5)$$

As a result of decomposition process of the full system, all elements belong to single-element modules. As an example of the concept of identifying the elements in complex system let's look at the decomposition of the system consisting of 18 elements (Table I).

In this example it is $l=5$ levels of decomposition. The number of modules at different levels is summarized in Table II.

TABLE I. NUMERICAL DETAILS OF SYSTEM DECOMPOSITION

Decomposition level i	Number of elements n_i	Number of multi-element modules n_i^*	Number of single-element modules d_i
1	4	1	3
2	7	2	5
3	10	7	3
4	6	5	1
5	2	2	0

Code assignment procedure requires, in the first step, to distribute all multiple modules, so that the structure at the lowest decomposition level consists only with non-disintegrable elements e_i , i.e. $E_{j_1 j_2 \dots j_k} = \{e_i\}$. Then e_i element is attributed to the index of the module, or code $j_1 j_2 \dots j_k$. It is written as: $e_i^{j_1 j_2 \dots j_k}$. Number k indicates that the element passes k levels of decomposition. After full decomposition of the

system, each element is given a unique code indicating its location in the functional structure. In the given example the e_7 has a code 3-2-1, so that we denote it: $e_7^{3,2,1}$, and e_8 has a code 3-2-2-1 so we denote it: $e_8^{3,2,2,1}$.

Let $K = \{k_1, k_2, \dots, k_n\}$ denotes the set of codes of decomposed system. The function code (6) is called element coding. This function is a bijection, and the inverse function assign element with e given code.

$$Code: E \rightarrow K \quad (6)$$

Presented coding system gives a complete knowledge of the structure of the system. Introduction set-theoretic formalism [1], taking into account the effect of degradation of the system on its declining utility can set a new direction of research on modelling the process of operation of the facility.

IV. TECHNICAL OBJECT AS A SYSTEM OF INDETERMINATE A PRIORI STRUCTURE (DYNAMIC STRUCTURE – “GROWING SYSTEM”)

In the real operation is often necessary to identify the components of the complex technical object while and-end are not supplied with catalogues describing object structure what saves an interest of technical services or protect intellectual property rights relating to technical solutions.

Description for the elements in the structure of the object becomes important if the complexity of the object is significant, i.e. entire system has hundreds or thousands of components or, on the other hand, if the knowledge of the position of element is not required for all items. The needs to identify the object element in the system appear each time when this element creates information that the end user may use. It means he has to undertake certain decision or actions, based on the event like: the element failed, item does not meet user requirements or spare part should be ordered.

In reliability field test of bucket wheel excavator SRs-2000 it were observed damages of 86 elements among 260 identified main modules on 3 decomposition levels [5]. Similar observation took place in the study of bucket loader L-220, where there are 259 elements failed among 3334 modules stored in the catalogue [6]. These examples show redundancy of the catalogues in relation to operational needs. The idea of dynamic catalogue would save time and memory of database and it can grow according to new events that appear in operation.

It is assumed that at the beginning of the service is not known object structure and its construction is "learnt" as a result of the occurrence of events related to the individual modules. Elements of the system create modules (subsystems) with common characteristics. Technical analysis allows for introduction of sub-elements for each structure as the appearance of information generated by them. The key to proper placement of elements in the structure is their characteristics and properties of components which have been already identified.

Element is the smallest part of the object i.e. not undergoing to further decomposition. Particularly, if the module is a single – piece module, it is called an element. Otherwise, we call it module. At the moment, if appears an event associated with a new module or component modification of the object structure is performed. Modification of the structure is done using three types of relations, in which

the new module can be recognised comparing to the known modules. These relations are: equivalence, subordination and the primacy relationship. Type of relationship for the new module to the already known modules is determined based on its properties.

TABLE II. AN EXAMPLE OF A COMPLETE SYSTEM DECOMPOSITION

Decomposition level	System elements that correspond to events describing operational process																		
	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9	e_{10}	e_{11}	e_{12}	e_{13}	e_{14}	e_{15}	e_{16}	e_{17}	e_{18}	
1	E_1^a	E_2			E_3				E_4										
2		E_{21}	E_{22}		E_{31}		E_{32}			E_{41}	E_{42}		E_{431}			E_{432}			
3			E_{221}	E_{222}	E_{311}	E_{312}	E_{321}	E_{322}			E_{421}	E_{422}							
4								E_{3221}	E_{3222}				E_{4311}	E_{4312}	E_{4313}	E_{4321}			
5																		E_{43221}	E_{43222}

^a. in bold, a single element module is represented

At the beginning of system structure creating, an object (system) is identified at the highest level of decomposition E. As soon as new event associated with the module e_1 appears, knowledge about the object is described as: $E = \{e_1\}$.

The first step of iterative structure modification requires analysis of the next event in the operation which may cause either a change in its structure as well as in the codes. The event may be related to:

1. the same element e_1 (just record the attribute for the element e_1)
2. new element e_2 which belongs to the same module (subsystem) as element e_1 on equivalent level of decomposition (siblings),
3. new element e_2 which belongs to the same module (subsystem) as element e_1 on subordinate level of decomposition (child),
4. new element e_2 which belongs to the same module (subsystem) as element e_1 on superior level of decomposition (parent),

According to the assumptions above, there is a description of the modification of the structure according to the scheme:

- ad 1. (as above): If you assume that the next event during the observation operating system is associated with existing in the structure module e_i then the structure of the system is not modified and information is saved as change of attribute of element e_1 : $E = \{e_i\}$.
- ad 2. (as above): Elements e_1 and e_2 form subsystems at the same level of decomposition, so that: $E = \{E_1, E_2\}; E_1 = \{e_1\}, E_2 = \{e_2\}$.

- ad 3. (as above): Element e_2 is an element of lower level of decomposition over e_1 , so that: $E = \{E_1, E_{11}\}; E_1 = \{e_1\}, E_{11} = \{e_2\}$,
- ad 4. (as above): Element e_2 is a super system element relatively to E_1 , so that: $E = \{E_1, E_{11}\}; E_1 = \{e_2\}, E_{11} = \{e_1\}$.

Table III shows an example of the development of the structure for the first seven elements. The table columns are related to the successive moments of event appearing in the system.

The example above shows code modification of originally single-element system requires algorithmic approach, taking into account the relationship between the elements (modules) already present in the structure and new elements: $e_i - R - E_j$.

The relation R shows the algorithm of code change for any system module:

- $R^=$ – equivalence position in the structure of the elements e_i and E_j (siblings) (7),

$$(e_i - R^= - E_j) \wedge (i = j_1 j_2 \dots j_k) \Rightarrow j = j_1 j_2 \dots (j_k + 1) \quad (7)$$

- R^- – subordinate position in the structure of elements e_i and E_j (child) (8)

$$(e_i - R^- - E_j) \wedge (i = j_1 j_2 \dots j_k) \Rightarrow j = j_1 j_2 \dots j_k j_{k+1} \quad (8)$$

- R^+ – primacy position in the structure of the elements e_i and E_j (parent) (9).

$$(e_i - R^+ - E_j) \wedge (i = j_1 j_2 \dots j_k) \Rightarrow j = i \wedge i = j_1 j_2 \dots j_{k-1} \quad (9)$$

In this case, the set of codes changes and after registration of the *i*-th element code set will take the form of *i*-element set: $K_i^* = \{k_1^*, k_2^*, \dots, k_i^*, \dots\}$, wherein $K_n^* = K$. Function *Code*: $E \rightarrow K_i^*$ maps the set of elements in the set of codes.

Iterative modification of the structure will be carried out until all modules are considered as single-element modules (elements).

TABLE III. MODIFICATION OF SYSTEM STRUCTURE WHILE CONSECUTIVE ELEMENTS ARE ADDED IN TIME

element	e_1	e_2 (siblings for E_1)	e_3 (child for E_1)	e_4 (child for E_{11})	e_5 (parent for E_2)	e_6 (siblings for E_1 and E_2)	e_7 (child for E_3)	e_i
<i>time of new element appearance</i>								
subsystem 1	$e_1 \Rightarrow E_1$		$e_3 \Rightarrow E_{11}$	$e_4 \Rightarrow E_{111}$...
subsystem 2		$e_2 \Rightarrow E_2$			$e_5 \Rightarrow E_2$...
subsystem 3						$e_6 \Rightarrow E_3$	$e_7 \Rightarrow E_{31}$...
subsystem ...		^{b)} $e_i \Rightarrow E_j$ – means that new element e_i is labelling E_j						...

SUMMARY

The paper presents two approaches to systematic identification of elements in a complex system. These approaches are particularly important at the stage of creating data bank in operational research. The first approach is the traditional one leading to the determination of structural or functional relationships for all elements that create an object. This approach ensures full knowledge of the construction of the system at every stage of research and analysis, however, requires considerable work and oversizing database of object structure. Dynamic approach presented in the paper modifies the structure of the object after every event appearance related to the "new" element, yet not existing in the structure. Database starts with one element, which is entire object and grows as new information about components come from the operation. This may limit the size of database because, as it is seen from observation, only 10-30% of the object components provide operational information.

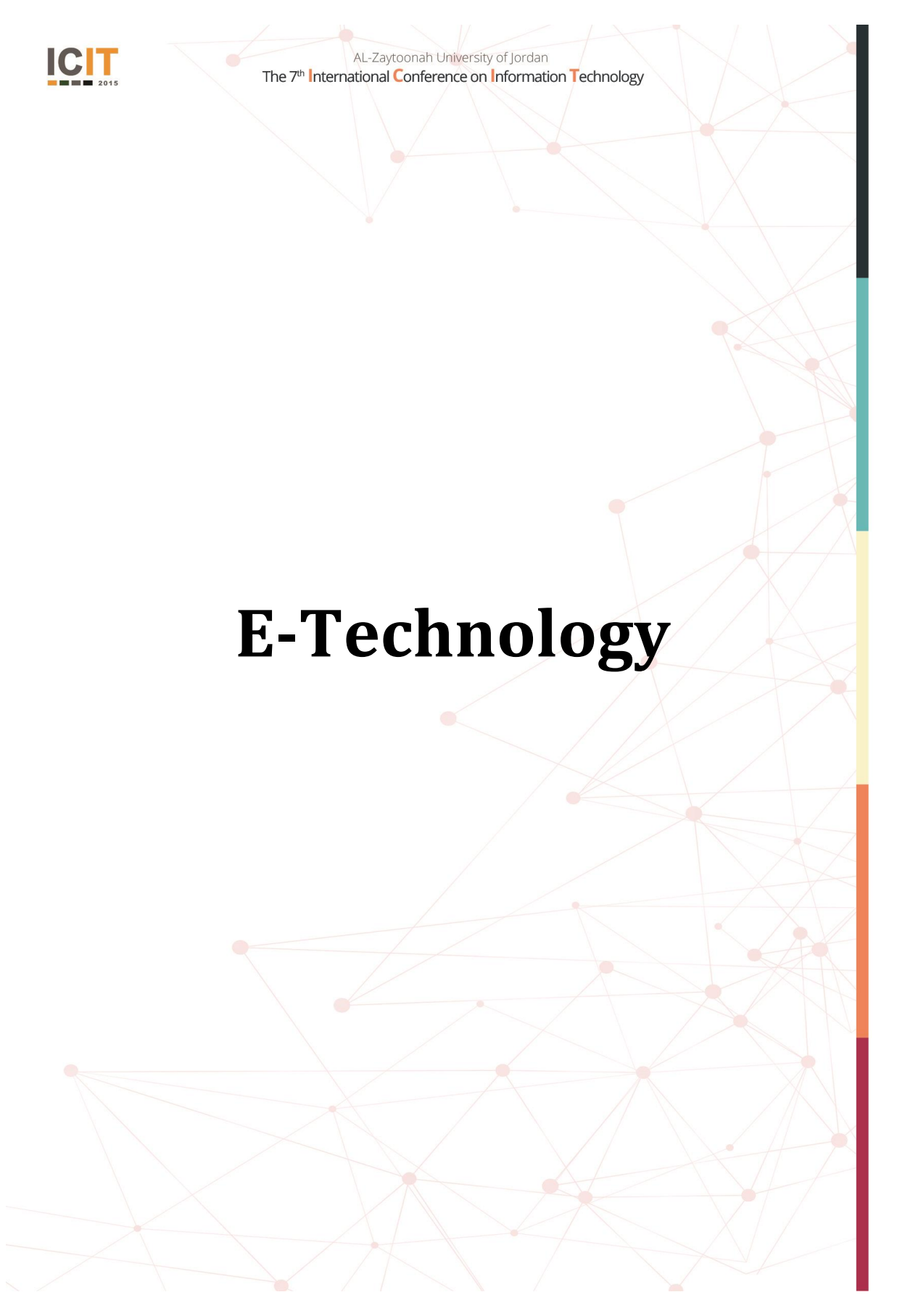
REFERECES

[1] K. Andrzejczak, "Set-theoretical approach to finite-state system reliability structure". Szczyrk: Proceedings of XXXIV Winter School of Reliability, 2006: 13-30.

[2] R.E. Barlow, F. "Proschan Mathematical theory of Reliability". New York: John Wiley and Sons Inc., 1965.
 [3] D. Bobrowski, "Models and methods of mathematical theory of reliability". Warsaw: Scientific and Technical Publishing, 1985.
 [4] E. Fidelis, S. Firkowicz, K. Grzesiak, K. Kołodziejski, K. Wisniewski, "Mathematical Foundations of reliability assessment". Warsaw: PWN, 1966.
 [5] A. Gołabek, M. Młyńczak, T. Nowakowski, "Computer system operational reliability assessment of open pit mine machines". Górnictwo odkrywkowe - Journal of Open Pit Mining 1992; 33 (3/4): 66-74.
 [6] A. Gołabek, M. Młyńczak, T. Nowakowski, "Assessment of operational reliability of L-220 loader". Overview of Mechanical Engineering - Mechanical Review 1990, 49 (18): 53-60.
 [7] A. Gołabek, "Procedures for testing and reliability assessment of machinery". Scientific Papers of the Institute of Machine Design and Operation, Wrocław University of Technology; Series of Monographs; 21 Wrocław: Ed. Wrocław University of Technology, 1992.
 [8] "Handbook of Performability Engineering". Ed.: K.B. Misra. London: Springer-Verlag, 2008.
 [9] A. Kossiakoff, W.N. Sweet, S.J. Seymour, S.M. Biemer, "Systems Engineering Principles and Practice". Hoboken, NJ: John Wiley & Sons, Inc., 2011.
 [10] "L220 Charger. Spare parts catalog". Factory Construction Equipment BUMAR-FADROMA. Wrocław: AGPOL, 1986.
 [11] A. Myers, "Complex System Reliability. Multichannel Systems with Imperfect Fault Coverage". London: Springer-Verlag, 2010.

- [12] "Reliability and maintenance of systems". Ed. Zamojski W. Wroclaw: Wroclaw University of Technology, 1981.
- [13] T. Nowakowski, "Problems with analyzing operational data uncertainty". Archives of Civil and Mechanical Engineering 2010; 10 (3): 95-109.
- [14] "Product Reliability. Specification and Performance". Ed. D.N.P.Murthy, M. Rausand, T. Østerås. London: Springer-Verlag, 2008.
- [15] W. Sadowski, "Fundamentals of general systems theory". Warsaw: PWN, 1978.
- [16] DJ Smith, "Reliability, Maintainability and Risk. Practical methods for engineers". Oxford: Butterworth-Heinemann, 2000.
- [17] W.R. Wessels, "Practical Reliability Engineering and Analysis for System Design and Life-Cycle Sustainment". Boca Raton: CRC Press Taylor & Francis Group, 2010.

E-Technology



NOVEL REVIEW OF ELECTRONIC GOVERNMENT STAGES AMONG DIFFERENT CONTINENTS

Munadil K. Faaeq

Department of Banking and Financial Science
College of Administrative and Financial Sciences
Cihan University
Erbil – Iraq
Monadela@yahoo.com

Alaa K Faieq

Baghdad Collage of Economic Sciences University
Baghdad – Iraq
alaa_dijun@yahoo.com

Mohammad M. Rasheed

IT Directorate
Ministry of Science and Technology
Baghdad - Iraq
mohmadmhr@yahoo.com

Thabit Hassan Thabit

Accounting Department
Cihan University
Erbil - Iraq
Thabit.acc@gmail.com

Abstract—This research explains Electronic Government (EG) stages around the world. Nowadays, there is in need for particular form to classify EG project stage in each country. EG project each country reaches a particular stage. These countries are trying to develop and enhance EG project by the available sources. The main objective is to systematic review of EG stages among many countries around the world. The finding of the current research is novel review of EG stage. In other words, this research review a significant and sophisticated EG stage among beneficiaries as following; (i)Readiness Stage, (ii)Initiative stage, (iii)Adoption Stage, (iv)Implementation stage, (v) and Developing and benchmark stage. Lastly, the findings will help governments spicily policy and strategic makers and practitioners to useful from the other countries experience and knows they current position.

Keywords :EG Stages, EG models, EG Services.

I. INTRODUCTION

Highlight Initially, Government is a public organization. It is an important part of broader governance systems. Its public agencies, set up by a society to help the beneficiaries getting the needed services and the information. This includes linking the society's development, related demands, and needs, then trying to collect them and implementing the right solutions to be more useful. Transparency is a necessary condition for government's responsibility vis-à-vis an oversight body [1]. Electronic-Government, or (EG), is a modern government that employs technology to transform and view its internal and external relationships. By applying the electronic technology in its operations, a government does not need to change its functions or its obligation to be useful, legitimate, transparent and responsible. In any case, these applications for the government will increase the expectations of its society about the performance of the government, in all regards, to a much higher level. In Middle East in general, there are scarcity of studies.

II. EG ADVANTAGES

There are many advantages that can be obtained from using EG. The important advantage of an EG is the increasing of the value of efficiency for the current systems. That will aid to save money and time for both the government and its beneficiaries. Furthermore, the EG facilitates the communication between governments and businesses. For example ,E-Procurement facilitates Government-to-Government communication; this example will permit smaller business to compete government agencies contracts as well as larger business. This will have the advantage of creating an open Electronic-Market and this E-Market will support the business to publish the business in the World Wide Web. In same time business and citizens can get information at a faster speed and it is possible at any time. In addition, which, when government used the electronic system will reduce the number of manpower. Truly, this would permit the process to be handled by lesser manpower and also to reduce the operations cost and time also[4].

III. EG SERVICES TYPES

The government provides services to many sides (beneficiaries) as following:

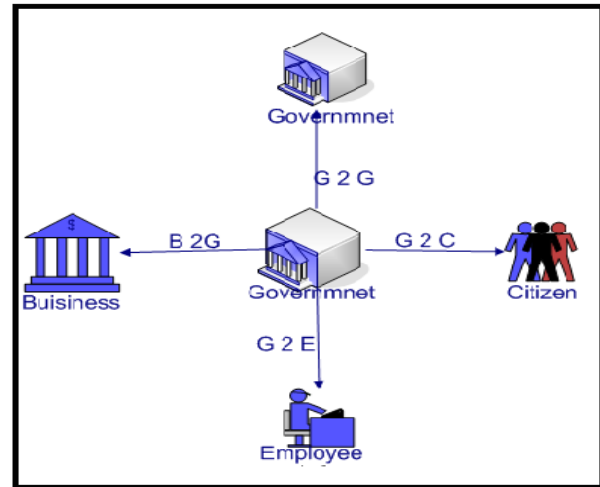


Fig 1. EG services types

- **Government to Citizen (G 2 C):** provide data and services from government agencies to citizen, including static document content and transactional systems such as tax payments, information inquiries, vehicle registration, permit processing, healthcare claims and services, social services delivery [5] increased the transparency leading and easy access to the government information that available at government website [6].
- **Government to Employees (G 2 E):** Provide streamlined services to government employees, including e-travel, e-training, expense reporting and reimbursement [5].
- **Government to Government (G 2 G):** Share data and transaction with other government organizations (agencies) to increase operational efficiencies, including grant management, loan processing, tax payment, processing, grant management [5].
- **Government to Business (G 2 B):** E-Procurement, applied in some countries like USA, Chile, Singapore and India, that aides to increase the doing of business with government [6]. Furthermore, it provides portal access to interoperate with businesses outside the agency, including purchasing portal, loan processing, and tax collection and processing [5].

IV. EG STAGES

EG started in July 2001 when the president of USA decided to start a strategy to come up with EG. Generally, EG around the world is trying to growing to be ready to face the challenges in developing, under developing and developed nations [5]. Figure 2 shows EG stages during the developing stage around the world during the period of time.

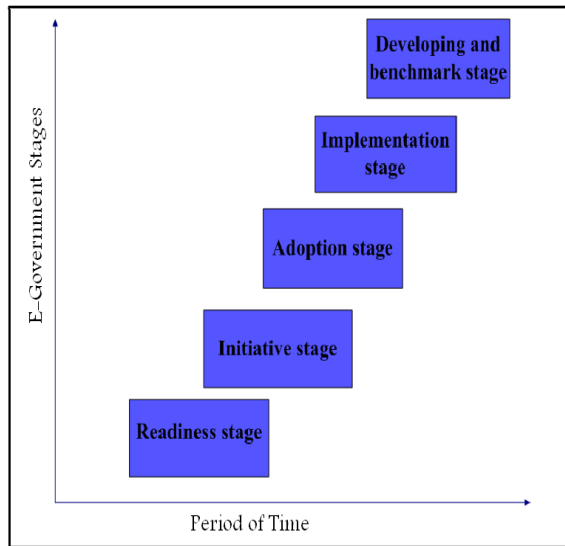


Fig 2. EG project stages.

A. Readiness Stage

Readiness is the first stage and the basic of launching EG around the world. This can be seen in Figure 2.

Actually, E-Readiness means build electronic infrastructure. At the same time, it should be integrated with what the country has from ICT's dealing with business communication as e-business, current ICT, government (EG) in the country, and nation too. A very strong basic of communication, both inside and outside the country, aid the trade and global investment (Harvard Business School). E-Ready is means that uses a computer in various places (schools, business, and government).

It means the use of a computer devise and get benefit from this technology in government, business, and school (McConnel International) [6]. Alsohybe, (2007) presents the readiness stage in some countries and focuses on Yamane. He used survey and interview to get an overview of the EG readiness in Yamane [7].

In fact, EG is a new government process via internet to increase the benefit from the business

sector. In this case, the innovation diffusion theory could be implemented and used [7].

Moreover, The government of Yemen is seeking an EG model that can aide in the implementation (as a advance stage) of EG and help the communication between the government and the stockholders in a short period[7].

The United Nations Global EG survey viewed that EG evolution must track national development, identify the difference in access and usetechnology, move on the waytoinclude theinformation of a society, and aid international comparisons [9]. Moreover the U.N. global survey aimed to examine governments' readiness to use EG application to improve and enhance the services thatare presented to citizens. The survey contributed to the development efforts of the member situation by focusing on whether EG impacts the socioeconomic support of the citizen lives. The survey provides a benchmark of a country's state of E-Readiness (a country's preparedness to integrate technology into society). The main goal of the survey were to provide and support an appraisal of the use of EG application to deliver and transport social services and to provide and give the comparative assessment of the readiness and ability of governments to engage citizens in e-Participation.

The U.N. Global EG Survey (2003) in Table 1showed the ranking of E-Readiness for some countries in the world regarding.

TABLE 1: GLOBAL EG READINESS RANKINGS 2003: TOP 23COUNTRIES [1]:

Country	Country
1. United States	13. Republic of Korea
2. Sweden	14. New Zealand
3. Australia	15. Iceland
4. Denmark	16. Estonia
5. United Kingdom	17. Ireland
6. Canada	18. Japan
7. Norway	19. France
8. Switzerland	20. Italy
9. Germany	21. Austria
10. Finland	22. Chile
11. Netherlands	23. Belgium
12. Singapore	

The report published by UN in 2005 mentions the transferring of EG to Electronic-Inclusion .Until now, Readiness is linked to the capability of three economic points' agents: individuals, firms, and government to capitalize on the use of ICT. In these days, the government performance can be measured by information system. Furthermore, the interesting perspective concentrated upon the interactions surrounded by ICT, individuals and

groups. In this side, the organizational and information system readiness, which mentions the implementation hiatus and transitional aid the correspondingly in Chau's research (1996), impact of user acceptance. Organizational Contingency Theories (OCT). However, it can provide some implications [10].

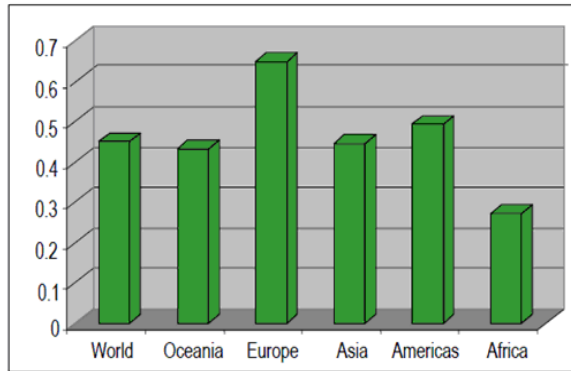


Fig 3. Region Average of EG readiness[11].

According to above discussion, there were many differences between those very important regions in terms of EG readiness, as following:- Europe regions (0.6490) having a clear improvement over the other regions, America got the second post (0.4936), followed Asia (0.4470), then Oceania (0.4338) and Africa (0.2739). Asia and Oceania were slimly below the world rate (0.4514), while Africa runs quite behind. Sweden got (0.9157) better than United States as the first region at that ranking. The Scandinavian countries group got the first three seats in the 2008 Survey, but Denmark (0.9134) took the second place and Norway (0.8921) third place. The United States (0.8644) reach the fourth place. At the ranking for the EG, European countries group got up 70% of the first 35 countries group at the same time the Asian countries group took 20 % of the first 35 countries. The infrastructure and connectivity has been investment got smile rate of failure in European countries group, quite especially in broadband infrastructure .It is important to report that is no one of country from African, Caribbean, Center or North American and Asia previously respectively took any seat at the top 35 countries at the 2008 E-Readiness ranking.

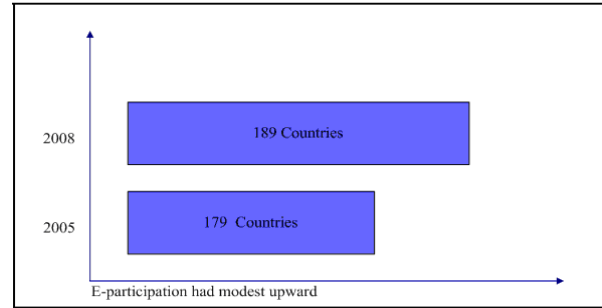


Fig 4. The E-Participation had modest upward from 179 countries in 2005 to 189 countries in 2008.

However, the US took the top place on the e-participation index. This was initially due to its effectiveness in e-information and also e-consultation, which aid its citizens to be more interesting toward government. It was directly followed during the Republic of Korea (0.9773), which performed truly fine in the e-consultation estimation. Denmark (0.9318) and France (0.9318) were took the third place[11].

TABLE 2. TOP 35 COUNTRIES IN THE 2008 EG READINESS[11].

Rank	Country	Rank	Country
1	Sweden	18	New Zealand
2	Denmark	19	Ireland
3	Norway	20	Spain
4	United States	21	Iceland
5	Netherlands	22	Germany
6	Republic of Korea	23	Singapore
7	Canada	24	Belgium
8	Australia	25	Czech Republic
9	France	26	Slovenia
10	United Kingdom	27	Italy
11	Japan	28	Lithuania
12	Switzerland	29	Malta
13	Estonia	30	Hungary
14	Luxembourg	31	Portugal
15	Finland	32	United Arab Emirates
16	Austria	33	Poland
17	Israel	34	Malaysia
		35	Cyprus

According to the Table 2 there are first 35 countries which were listed among readiness stage in the whole world.

TABLE 3: REGIONAL EG READINESS RANKING[11].

Region	2008	2005	Region	2008	2005
Africa			Americas		
Central Africa	0.2530	0.2397	Caribbean	0.4480	0.4282
Eastern Africa	0.2879	0.2836	Central America	0.4604	0.4255
Northern Africa	0.3403	0.3098	North America	0.8408	0.8744
Southern Africa	0.3893	0.3896	South America	0.5072	0.4901
West Africa	0.2110	0.1930			
Asia			Europe		
Central Asia	0.3881	0.4173	Eastern Europe	0.5689	0.5556
Eastern Asia	0.6443	0.6392	Northern Europe	0.7721	0.7751
Southern Asia	0.3395	0.3126	Southern Europe	0.5642	0.4654
South-Eastern Asia	0.4290	0.4388	Western Europe	0.7329	0.6248
Western Asia	0.4857	0.4384			
Oceania	0.4338	0.2888			
World Average	0.4514	0.4267			

In this report, the survey helps to develop the member's countries during focusing on socio-economic up life and how these factors influence EG. This survey highlights the country's state about the E-Readiness level[14].

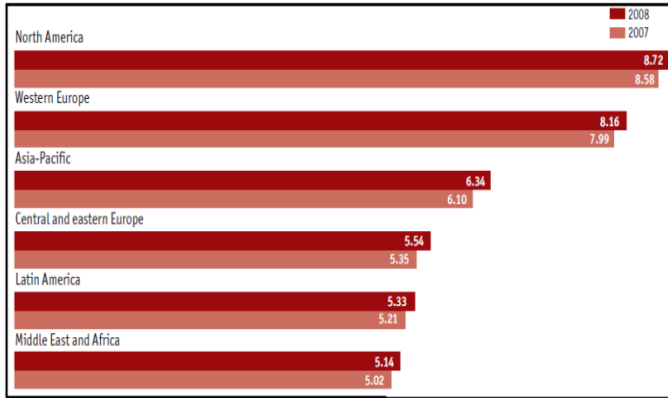


Fig. 3: The six world region scores for Economist Intelligence Unit-Readiness rankings,

2008.

According to the Figure 3, there are six major regions in the world. As the figure shows, there are big differences between them. At the same time, there is difference in the same region under a short period of time (two years). That means the regions push up itself to be better. Also, there are differences in that scour as the figure mentions it. At Table 4 shows the compare E-Readiness ranking from 2007 to 2008 for first 70 countries round the world, the first ten countries in 2008 are as following sequence: United States, Hong Kong, Sweden, Australia, Denmark, Singapore, Netherlands, United Kingdom, Switzerland and Austria. Furthermore there are 4 Arabic countries that occupy deferent position as: United Arab Emirates, Saudi Arabia, Jordan, and Egypt. The reasons behind shifting many countries return to policy and communication network.

TABLE 4. ECONOMIST INTELLIGENCE UNIT E-READINESS RANKINGS, 2008 [13].

Economist Intelligence Unit e-readiness rankings, 2008									
2008 e-readiness rank (of 70)	2007 rank	Country	2008 e-readiness score (of 10)	2007 score	2008 e-readiness rank (of 70)	2007 rank	Country	2008 e-readiness score (of 10)	2007 score
1	2	United States	8.95	8.85	36	39	Slovakia	6.06	5.84
2	4	Hong Kong	8.91	8.72	37	37	Latvia	6.03	5.88
3	2	Sweden	8.85	8.85	38	41	Lithuania	6.03	5.78
4	9	Australia	8.83	8.46	39	35	South Africa	5.95	6.10
5	1	Denmark	8.83	8.88	40	38	Mexico	5.88	5.86
6	6	Singapore	8.74	8.60	41	40	Poland	5.83	5.80
7	8	Netherlands	8.74	8.50	42	43	Brazil	5.65	5.45
8	7	United Kingdom	8.68	8.59	43	42	Turkey	5.64	5.61
9	5	Switzerland	8.67	8.61	44	44	Argentina	5.56	5.40
10	11	Austria	8.63	8.39	45	45	Romania	5.46	5.32
11	12	Norway	8.60	8.35	46	46	Saudi Arabia	5.23	5.05
12	13	Canada	8.49	8.30	47	49	Thailand	5.22	4.91
13	10	Finland	8.42	8.43	48	48	Bulgaria	5.19	5.01
14	19	Germany	8.39	8.00	49	46	Jamaica	5.17	5.05
15	16	South Korea	8.34	8.08	50	—	Trinidad & Tobago*	5.07	—
16	14	New Zealand	8.28	8.19	51	51	Peru	5.07	4.83
17	15	Bermuda	8.22	8.15	52	50	Venezuela	5.06	4.89
18	18	Japan	8.08	8.01	53	52	Jordan	5.03	4.77
19	17	Taiwan	8.05	8.05	54	54	India	4.96	4.66
20	20	Belgium	8.04	7.90	55	54	Philippines	4.90	4.66
21	21	Ireland	8.03	7.86	56	56	China	4.85	4.43
22	22	France	7.92	7.77	57	58	Egypt	4.81	4.26
23	24	Malta	7.78	7.56	58	53	Colombia	4.71	4.69
24	23	Israel	7.61	7.58	59	57	Russia	4.42	4.27
25	25	Italy	7.55	7.45	60	61	Sri Lanka	4.35	3.93
26	26	Spain	7.46	7.29	61	60	Ukraine	4.31	4.02
27	27	Portugal	7.38	7.14	62	62	Nigeria	4.25	3.92
28	28	Estonia	7.10	6.84	63	59	Ecuador	4.17	4.12
29	29	Slovenia	6.93	6.66	64	63	Pakistan	4.10	3.79
30	32	Greece	6.72	6.31	65	65	Vietnam	4.03	3.73
31	31	Czech Republic	6.68	6.32	66	64	Kazakhstan	3.89	3.78
32	30	Chile	6.57	6.47	67	66	Algeria	3.61	3.63
33	34	Hungary	6.30	6.16	68	67	Indonesia	3.59	3.39
34	36	Malaysia	6.16	5.97	69	68	Azerbaijan	3.29	3.26
35	33	United Arab Emirates	6.09	6.22	70	69	Iran	3.18	3.08

* New to the annual rankings in 2008. Note: A four-decimal score is used to determine each country's rank. Source: Economist Intelligence Unit, 2008.

B. Initiative stage

Initiative means "the power or ability to begin or to follow through energetically with a plan or task; enterprise and determination"[20].Subsequence, this is the second stage in this research. Furthermore, there is one theory named Initiative theory.

Furthermore, there are four variables very important framework related to Roger as following:

The innovation, Communication Channels, Social System, and Time.However that frame is used and applied at various areaslike: - public and privet sectors. EG is very big area and get interesting from many researchers and authors, and there are many avenues in this area. It is not explored yet.EG is one of the very important project, in public sector needs to comprehensive assessment in periodicity way. Evaluation of IT in general and in specific way in IS Acceptance is very important.

Regarding to this study, the authors show ex-pot from work for EG project. They mention three dimensionalframeworks for EG initiatives. During three domains of EG maturity levels, stockholder and Assessment levels and how that influence on EG initiative. Currently the range of the government agencies increasing to usage of IS at dailytask.There are scarcity of the information on the quality and efficiency of EG Initiative that lead tothe weakness of the evaluation EG quality[14].Recently, the success rates of EG projects are estimated to be less 15% [14]. The researchers and the studies that related on EG area are increased [14].

In same filed, there is a study that discusses and analysis of EG intuitive throughout evaluating the framework. This study mentions that there is no standardized measure for evaluating the impact of EG initiatives. However the survey improves that there is positive influence on the national and universal In case EG means are positively managed.EG policies are linked with different domain like economic, social and country infrastructures. There are many platform builders –new EG initiatives (Japan, Brazil, and Malaysia). In fact, this project focuses on economic filed that linked to EGinitiatives.

There are very important notes that should be mentioned here. There are some projects having short terms and the benefits from them can appear at short time also[14] for exampleSTOPE project[14].

C. Adoption Stage

The free dictionary presented adoption as a " act of accepting with approval; favorable reception"[21]. Generally, in US and at the three Latin American countries (Argentina, Brazil and Mexico) 2008. There is a very important reference to third nations and the researchers. Generally, government agencies should control the change management operation and the ability to applicable on the adoption on EG projects [15].

This study is divided into two parts. The first part of this paper reviewed the conceptual framework to test the development and the services of EG. The second part discussed the findings and highlighted on different nations on each countries as model for successful frame for positive improvement as well as the EG in a non –industrialized and also in developing countries[15].

D. Implementation stage

Implementation defined as act of accomplishing some aim or executing some order [22]. Currently, the implementation is the fourth stage in the EG project. At the implementation stage in Spain the government tarried to linked all Spanish city to enhance and gave the opportunity to various agencies to provide the services to the stockholders (26) at this stage the Singaporeans EG tried to engage the stockholders to practices of EG systems[17].The requirement of this stage is to use model that is aide the agencies to enhance the efficiency of the service to the stockholders. This study reviewed the macro perspective of different action the implementation of EG during analysis of different EG related initiatives under taken by the Singapore government. The analysis operation guide in the implementation EG likes:

- ✓ IC (Information Center).
- ✓ ICT infrastructure.
- ✓ EG infrastructure.
- ✓ EG promotion[17].

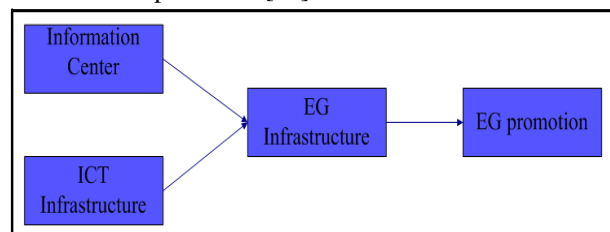


Fig. 4.These four items were very important to EG implementation framework [17]

This study expects the specific frame may be it used like a tool to organize and arrange EG project or to used and done strategy of EG implementation. EG strategy and plan, process proceed from exception of environment that has ability to dealing with internet technology and communication tools with public sector.

The Spain's low agree that exception to aide each city councils to be able to provide the serves electronically [18].

E. Developing and benchmark stage

The motoring, evaluating and benchmarking EG very important to policy and dictions makers to better evaluation criteria for their decision and developing these countries [14]. This study emphasizes conception and analysis of EG project [14].

There is interesting research viewed the influence of public participants on EG in three places Sweden, Bygga and Villa. The researcher was dealing with 16 organizations at different levels of society, including Education, Government, and Industry to aide and develop an innovative. The electronic portal for private construction industry. Therefore, the research aimed to view the challenges and how they can be overcome. [19].

V. Findings

This study could be the reference and guide for the researchers, students and how they interested in EG because, this research viewed positions and challenges for many countries around the world.

This study shows sequence of EG stages by selected countries to view each country where it reach now up on available documents (studies, reports).

There is signal that certifies that we are going at the right track. We can see Brazil was shown at two stages but in deferent time,

Furthermore, Brazil viewed at Initiative stage in 2007

But after one year, Brazil comes up again at different stage in Adoption stage (in advance) at 2008. Those mean we going at the right way in this study.

Furthermore, we present many problems and issues related to the stages that countries were reached like:

Pinpointing the reasons and the means to increase the success rate of EG project. Also, Exploring framework to aid organize and coordinate different EG projects. Come up with scientific research related to acceptance technology, Shown three dimensional frameworks related to EG project.

Moreover, the overview of the rate of EG project for many countries around the world for example, Estonia, Swedish, United States, UK, Canada, Malaysia, Egypt and Yemen, etc.

These challenges and issues are the reason for why these countries are late. From this overview we can see the deferent of positions among the countries and at same time same country in deferent times. Same issue for region to know the problems and processing it as fast as they can. Different developing nations should pass likewise a problem by this study.

VI. Future works

For future work, any country try to enhance information system specially EG project. Developing countries' society needs more applications to be applied. We strongly recommend launching Electronic-Census and publish information kiosks in different region for many reason like security issue for the citizen and also this application will decrees the effort, corruption, time and money to the employees in the first hand and citizen from the anther hand.

VII. Conclusion

We can get benefit from this overview to know what the issues that countries have to face and how can pass it by governments to enhance EG project. In brief, governments can use this study to know the position map for many countries to get the benefit from it.

Furthermore government can study them (countries was viewed in this study) to enhance EG project. It is useful for researchers and government for several reasons. First, we were preview important models related to three countries these countries took good positions at EG ranking in the world. Second, this research identifies the essential factors that might lead to the adoption and implementation of a successful EG plan. Finally, it views the challenges that the country may face.

References

- [1] United Nations, "World Public Sector Report 2003," New York 2003.
- [2] A. K. Faieq, M. K. Faieq, and H. A. Hambali, "A requirement model for Baghdad University's E-magazine System for Scientific Research " in *EEE'09 - The 2009 International Conference on e-Learning, e-Business, Enterprise Information Systems, and e-Government Las Vegas, USA, 2009*.
- [3] M. K. Faieq, N. A. Ismail, W. R. S. Osman, A. K. Faieq, and M. Aleqab, "Application of TAM in the readiness for E-Government services," in *12th IBIMA conference on Creating Global Economies through Innovation and Knowledge Management Kuala Lumpur, Malaysia, 2009*.
- [4] "MSG's ITLG blog," vol. 2009, 2009.
- [5] A. Doko, D. Adarnalee, and Y. A. Salem, "A Workflow Management System for E-Government " in *Information Technology Engineering*, vol. Graduation Damascus: Damascus University.
- [6] J. Satyanarayana, *e-Government*. New Delhi, 2006, p.140.
- [7] N. T. Alsohybe, "The implementation of e-government in the republic of yemen: an empirical evaluation of the technical and organizational readiness," Capella University, 2007.
- [8] Y. S. Wang, Y. W. Shih, "Why do people use information kiosks? A validation of the Unified Theory of Acceptance and Use of Technology," *Government Information Quarterly*, vol. 26, pp. 158–165, 2009.
- [9] United Nations, "UN Global E-government Readiness Report 2005," 2005.
- [10] H. Sun and P. Zhang, "A Methodological Analysis of User Technology Acceptance," in *37th Hawaii International Conference on System Sciences, Hawaii, 2004*.
- [11] United Nations, "United Nations e-Government Survey 2008," New York 2008.
- [12] R. Kumar, M. L. Best, "Impact and Sustainability of E-Government Services in Developing Countries: Lessons Learned from Tamil, India," pp. 1–12, 2005.
- [13] IBM Institute for Business Value, "E-readiness rankings 2008 Maintaining momentum," 2008.
- [14] J. Esteves and R. C. Joseph, "A comprehensive framework for the assessment of eGovernment projects," *Government Information Quarterly*, 2007.
- [15] T. Y. Lau, M. Aboulhosen, C. Lin, and D. J. Atkin, "Adoption of e-government in three Latin American countries: Argentina, Brazil and Mexico," *Telecommunications Policy*, pp. 88–100, 2008.
- [16] S. C. Y. Luk, "The Impact of E-government in Greater China: Case Studies of Hong Kong, Taiwan, and Singapore1," *17th Biennial Conference of the Asian Studies Association of Australia in Melbourne 1-3 July 2008*, pp. 1–23, 2008.
- [17] C. M. L. Chan, Y. Lau, and S. L. Pan, "E-government implementation: A macro analysis of Singapore's e-government initiatives," *Government Information Quarterly*, vol. 25, pp. 239–255, 2008.
- [18] S. de Juana- Espinosa, J. Valdés-Conca, E. Manresa-Marhuenda, and L. García-Felonés, "E-Government Implementation in Spain: the Case of the City of Benidorm," *Innovation and Knowledge Management in Business Globalization: Theory & Practice*, pp. 1125–1130.
- [19] I. Ruuska and R. Teigland, "Ensuring project success through collective competence and creative conflict in public-private partnerships – A case study of Bygga Villa, a Swedish triple helix e-government initiative," *International Journal of Project Management*, 2008.
- [20] TheFreeDictionary <http://www.thefreedictionary.com/initiative>. Retrieved February 15, 2012.
- [21] TheFreeDictionary <http://www.thefreedictionary.com/Adoption>. Retrieved February 15, 2012
- [23] iraq electronic government portal <http://www.egov.gov.iq/egoviraq/index.jsp?sid=1&id=263&pid=250>. Retrieved February 16, 2012

Predictors of Mobile Learning Adoption

The Case of Universiti Teknologi MARA

Mohamad Noorman Masrek

Accounting Research Institute & Faculty of Information Management
Universiti Teknologi MARA
Shah Alam Selangor, Malaysia
mnoorman@salam.uitm.edu.my

Abstract—Despite the availability of studies on mobile learning adoption, its theoretical foundations have not yet matured. In addition, studies on mobile learning adoption in the context of Malaysia is also still very limited. Against this concern, a study was undertaken with the aim of investigating factors that could influence mobile learning adopting. Drawing upon The Unified Theory of Acceptance and Use of Technology (UTAUT) and two other variables which are perceived playfulness and self management of learning, an empirical based framework was developed to identify predictors of mobile learning. Employing survey research method involving 282 respondents from Universiti Teknologi MARA, the results showed that performance expectancy, effort expectancy, social factors, facilitating conditions, perceived playfulness and self management of learning are strong determinants of intention to adopt mobile learning. The present study provides both a theoretical and practical contributions to understanding the predictors of intention to adopt mobile learning and should be of interest to both researchers and practitioners.

Keywords—mobile learning, predictors, survey, structural equation modelling

I. INTRODUCTION

The revolution brought about by mobile technologies has resulted to the emergence of mobile learning, which is the extension or prolongation of e-learning. [30] described mobile-learning as a learning process which takes the advantages of mobile devices, ubiquitous communications technology and intelligent user interfaces. In universities, mobile learning helps educational institutions to enhance the accessibility, interoperability and reusability of educational resources, and also to improve flexibility and interactivity of learning behaviours at convenient times and places [2]. For learners in general, mobile learning facilitates the use of previously unproductive time, enables learning behaviours regardless of time and place; and brings about great possibilities for personalized, customized and context-aware learning support services [32].

Despite the availability of studies on mobile learning, its theoretical foundations have not yet matured [25]. According to [24], regardless of the high degree of insertion of mobile devices in current society, the mere availability of technology itself does not guarantee that its potential will be used for learning or accepted by all evenly. [7] also stressed that, the understanding of the adoption of mobile technologies in educational environments is still incipient and in particular, questions about how to promote the acceptance of mobile learning by users are still largely unresolved. Against this background, a study was conducted with the following objectives: (i) to identify factors that influence mobile learning adoption among students in higher learning institution in Malaysia, and (ii) to ascertain whether the following factors

influence intention to adopt mobile learning: performance expectancy, effort expectancy, social factors, facilitating conditions, perceived playfulness and self management of learning.

The rest of this paper is structured as follows. Firstly, this study presents the literature review on mobile learning, focusing on its concepts and related theories, model or framework used by researchers for studying mobile learning adoption. Secondly, it describes the research method, giving the details on measurement, population, sampling and data collection. Thirdly, it presents the results based on the data analysis. Fourthly, it presents the discussion on the research findings. Finally, it draws the conclusion of the study.

II. LITERATURE REVIEW

Mobile learning is defined as “handheld technologies, together with wireless and mobile phone networks, to facilitate, support, enhance and extend the reach of teaching and learning” [5]. On the other hand, Geddes defined mobile learning as “acquisition of any knowledge and skill through the use of mobile technology, anywhere, anytime that results in an alteration in behavior” [29]. [19] explained that mobile learning is highly situated, personal, collaborative and long term. Mobile learning is also considered as truly promoting learner-centred learning because of the following features: (i) portability - the small size and weight of mobile devices means they can be carried everywhere and help learning occur at anywhere and anytime; (ii) connectivity - providing learners with connections to other learning such as through other people, devices or networks; (iii) interactivity - mobile devices are potential tools for enhancing a cooperative

learning environment; (iv) context sensitivity - mobile devices enable learning to take place which can make greater use of a person's; (v) immediate context and surroundings; (vi) lifelong - mobile content consumption is continuous, there is no beginning, middle or end; (vii) individuality – learning can be customised and based on previous learning experiences [5],[19]. The advantages of mobile learning are (i) just-enough learning – highly applied, easily digestible learning for increasingly busy executives; (ii) just-in-time learning – convenient, flexible and relevant learning at the exact moment learning is required; (iii) just-for-me learning – learner-driven learning in a suitable format; (iv) cost-saving – mobile learning can be cost effective and using a learner's own mobile device eliminates technological barriers to accessing learning [19].

III. THEORETICAL FRAMEWORK

Since the dawn of mobile learning, researchers have studied factors that influence its adoption. Theories, models or framework such as Theory of Reasoned Action (TRA) [20]; Social Cognitive Theory (SCT) [1]; Technology Acceptance Model (TAM) [8]; Theory of Planned Behavior (TPB) [13]; Model of PC Utilization (MPCU) [38]; Innovation Diffusion Theory (IDT) [6]; Combined TAM and TPB [37]; and The Unified Theory of Acceptance and Use of Technology (UTAUT) [40] have been referred and adapted by researchers to investigate the mobile learning adoption. Among the various theories and models, UTAUT is found to be the most adopted or referred in the context of mobile learning. According to Masrom & Hussein, UTAUT could explain up to 70% of technology acceptance behavior [23]. UTAUT suggests that four key constructs which are, performance expectancy, effort expectancy, social factors and facilitating conditions have a direct influence on intention to adopt technology. Studies on mobile learning had empirically proof the contribution of these four constructs. Besides these four constructs, researchers have also explored the role of perceived playfulness and self management of learning. Drawing upon this premise, the present study will investigate the adoption of mobile learning based on the framework shown in Figure 1.

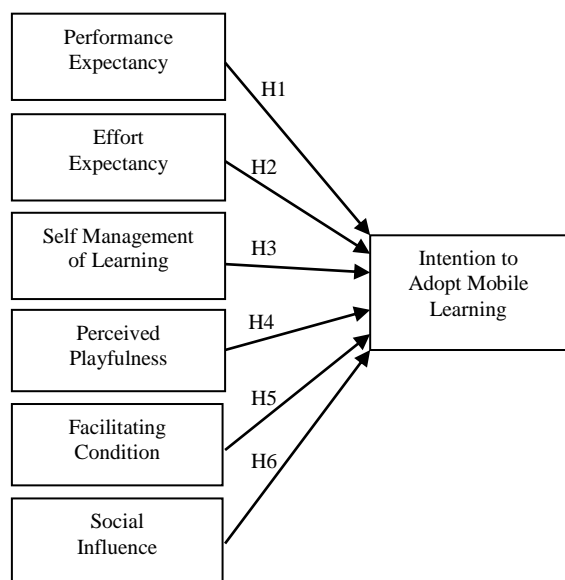


Fig. 1. Theoretical framework

A. Intention to Adopt Mobile Learning

Intention to adopt mobile learning is defined as “the person’s subjective probability that he or she will perform the behavior in question” [40]. In the context mobile learning adoption, various factors have been identified as predictors of intention to adopt. Momani & Abualkishik compiled a comprehensive list of factors which are perceived mobility; perceived ease of use; perceived usefulness; alignment value; intrinsic value; utility value; self-management of learning; comfort with mobile learning; perceived trust; performance expectancy; effort expectancy; social influence; perceived playfulness; relative advantage; facilitating condition; previous experience; resistance; importance of the course; integration of the technology into course assessment; lecturer modelling of the course; available tools; lecturer’s feedback; mobile device and software; perceived innovativeness; perceived ICT anxiety; perceived self efficacy; compatibility; complexity; trialability; observability; image; voluntariness; cost and perceived credibility [9]. Upon further scrutiny, Momani & Abualkishik identified eight most frequently examined factors which are performance expectancy; effort expectancy; self management of learning; social influence; facilitating conditions; perceived playfulness, perceived cost and previous experience [9].

B. Performance Expectancy

Performance expectancy, which is described by [40], refers to “the degree to which an individual believes that using the system will help him or her to attain gains in job performance”. [41] stated that adapting performance expectancy to mobile learning suggests that users will find mobile learning useful because it enables learners to accomplish learning activities more quickly, effectively and flexibly. Their study involving 330 respondents in Taiwan discovered that performance expectancy was the strongest predictor of intention to adopt mobile learning. Recent study in the context of Malaysia by Jambulingan also showed consistent finding [22]. Other studies that had also discovered similar findings are from [3], [10] [28], and [33]. Against this background, this study hypothesizes: *H1: Performance expectancy significantly affects intention to adopt mobile learning.*

C. Effort Expectancy

According to Venkatesh *et. al* effort expectancy is "the degree of ease associated with the use of the system" [40]. In the context of mobile learning, effort expectancy is about an individual's expectations of using mobile learning without

much effort. The easier the mobile learning applications can be accessed by the user, the more is the intention to adopt it. Studies across different countries showed mixed results on the influence of effort expectancy on intention to adopt mobile learning. While [22] did not find any support, others such as [3], [10], [17], [28], and [41] found positive relationship between the effort expectancy and intention to adopt mobile learning. Based on the aforementioned premise, this study posits that: *H2: Effort expectancy significantly affects intention to adopt mobile learning.*

D. Self Management of Learning

Self management of learning is defined as the extent to which an individual feels he or she is self-disciplined and can engage in autonomous learning [34]. Indeed, the need for self-direction, or self-management of learning, runs clearly throughout the distance education and resource-based flexible learning literature [31], [34], [42]. Since mobile learning can be considered as a kind of e-learning via mobile devices, it is expected that a person’s level of self-management of learning will have a positive influence on his or her behavioral intention to adopt mobile learning. Previous studies done by Wang *et. al* in the context of mobile learning found that self management of learning positively predicts intention to adopt mobile learning [41]. To this effect, this study hypothesizes that: *H3 – Self-management of learning significantly affects individual intention to adopt mobile learning.*

E. Perceived Playfulness

Perceived playfulness is considered one of the critical factors that could potentially affect learning engagement with the utilisation of new teaching innovations and technology [36]. Agarwal & Karahanna stated that perceived playfulness will provide intrinsic motivation when individuals become completely absorbed in a technology [26]. An intrinsic motivator refers to individual’s performance or engagement in an activity due to his or her interest in the activity [28]. Previous studies have also showed that the use of IT is influenced by perceived playfulness-related constructs [14], [26]. The reason is because individuals who experience pleasure or enjoyment from using an information system are more likely to intend to use it extensively than those who do not [21], [39]. Taken the above together, this study hypothesizes: *H4 - Perceived playfulness significantly affects individual intention to adopt mobile learning.*

F. Facilitating Conditions

Facilitating condition is defined as "the degree to which an individual believes that an organizational and technical infrastructure exists to support the use of the system" [40]. Acceptance of any new technology is highly dependent upon the supporting conditions or environment [28]. In the context of mobile learning, these facilitating conditions can appear in

the form such as resources, knowledge, Internet speed, and support personnel [28]. Studies reported by [17], [28] showed that facilitating condition is a significant predictor of mobile learning adoption. Given this background, this study postulates that: *H5: Facilitating conditions significantly affects intention to adopt mobile learning.*

G. Social Influence

Social influence is defined as "the degree to which an individual perceives that others believe he or she should use the new system" [40]. Thompson *et. al* called social influence as social factors and defined it as "the individual's internalization of the reference groups' subjective culture, and specific interpersonal agreements that the individual has made with others, in specific social situations" [38]. Kelman defined social influence with three different forms in his theory (i) compliance: when an individual accepts influence because he hopes to achieve a favourable reaction from another person or group (social approval/disapproval from others) (ii) identification: when an individual accepts influence because he wants to establish or maintain a satisfying self defining relationship with others; (iii) internalization: when an individual accepts influence because it is congruent with her value system [12]. De Silva *et al.* found that social influence in mobile adoption appeared in two modes: one that exerts pressure on individuals to adopt, and another that helps to generate benefits via social networks that are tied in with economic and business networks [11]. Consequently, grounded in UTAUT and justified by previous studies [3], [10], [17], [28], and [41], the following hypothesis is put forth: *H6: social influence significantly affects intention to adopt mobile learning.*

IV. RESEARCH METHODOLOGY

The study used survey method with questionnaire as the instrument for data collection. The questionnaire was based on the instruments used by previous studies. Perceptual measures in the form of statements were used for measuring each variable with a corresponding Likert scale anchored as 1 for “Strongly Disagree”; 2 for “Disagree”; 3 for “Neither Agree Nor Disagree”; 4 for “Agree” and 5 for “Strongly Agree”. The questionnaire was pre-tested with several experts and prospective respondents. Subsequently, it was pilot tested with 30 students. The results of the pilot test is illustrated in Table I showed that the Cronbach Alpha for all variables were well above 0.7, indicating that the questionnaire was acceptably reliable.

TABLE I. SOURCES OF MEASUREMENT

Variable	No of items	Sources of measurement	Cronbach Alpha of pilot test
Intention to adopt mobile learning	4	[28], [41]	0.76
Performance	4	[28], [41]	0.723

Variable	No of items	Sources of measurement	Cronbach Alpha of pilot test
expectancy			
Effort expectancy	4	[28], [41]	0.811
Self management of learning	3	[28], [41]	0.746
Perceived playfulness	3	[28], [41]	0.722
Facilitating conditions	3	[28], [41]	0.707
Social norms	3	[28], [41]	0.812

The population of the study was students enrolled to the bachelors degree in the Faculty of Information Management, Universiti Teknologi MARA, Malaysia. Using the simple random sampling technique, a total of 350 questionnaires were sent to the targeted students. The duration of data collection was one month and after the period was over, a total of 302 questionnaires were returned. However, 20 were found to be unusable for further analysis as they were incomplete. The remaining 282 were analyzed using IBM SPSS and AMOS version 20. The statistical analyses carried out were frequency analysis; descriptive analysis focusing on median, standard deviation, variance and testing normality of distribution; an exploratory factor analysis (EFA) for assessing unidimensionality; confirmatory factor analysis (CFA) for assessing convergent validity and discriminant validity; and structural equation modelling (SEM) or structural model for testing the established hypotheses.

V. FINDINGS

Table II showcases the demographic profile of the respondents. Out of 282 respondents, 73.8% were female while the remaining 26.2% were male. In terms of semester of study, the majority indicated to be in semester three (22.7%) while the minority were from semester six (12.4%). With regard to program registered, the majority of the respondents was doing BSc Information Management Systems (27.0%) and followed by BSc Library Science (25.5%).

TABLE II. DEMOGRAPHIC PROFILES

Variable		Frequency	Percent
Gender	Male	74	26.2
	Female	208	73.8
Semester	1	36	12.8
	2	54	19.1
	3	64	22.7
	4	42	14.9
	5	51	18.1
	6	35	12.4
Programs	BSc Library Science	72	25.5
	BSc Information Management Systems	76	27.0
	BSc Records Management	67	23.8
	BSc Resource Centre Management	67	23.8

In order to identify whether the data is experiencing common method bias, Harman's single factor test was executed. All items from all constructs under study were

entered for analysis and constrained to only single factor. The results showed that the single factor explained only 26.29%, less than the benchmark value of 50% of the total variance, implying that the collected data is free from the problem of common method variance. Normality testing on univariate and multivariate was also accessed upon the data. To test for univariate normality the skewness and kurtosis of each observed variable was assessed. The skewness and kurtosis requirements fulfilled the benchmark values suggested by [27] which are 3 and 10 respectively. To assess multivariate normality, [18] suggested that the Mardia's coefficient should be less than $p(p+2)$, where p is the number of observed variables. This study has 24 observed variable, so $24(24+2) = 624$. The Amos output for Mardia's coefficient is 68.56, which is less than 624, hence multivariate normality can be assumed.

Following [15], this study used factor loadings, composite reliability (CR) and average variance extracted (AVE) to measure the convergent validity. The recommended value of factor loadings should be above the value of 0.6 and as shown in Table 3, all the factor loadings met this requirement [35]. In terms of composite reliability, all the scores are well above the cut off value of 0.7 as recommended by [16]. [4] The literature suggests that the acceptable level of AVE should be more than 0.5 which is also fulfilled in this study as illustrated in Table 3. Accordingly, the study also assessed discriminant validity and the results are presented in Table III. [4] also suggested that AVE can also be used to assess discriminant validity by comparing its square root against the correlation values between the variables and all other variables. As displayed in Table IV, the square root of the AVE values is well above the correlation values, hence suggesting discriminant validity requirement is fully complied. The AMOS output of the measurement model is shown in Figure 2.

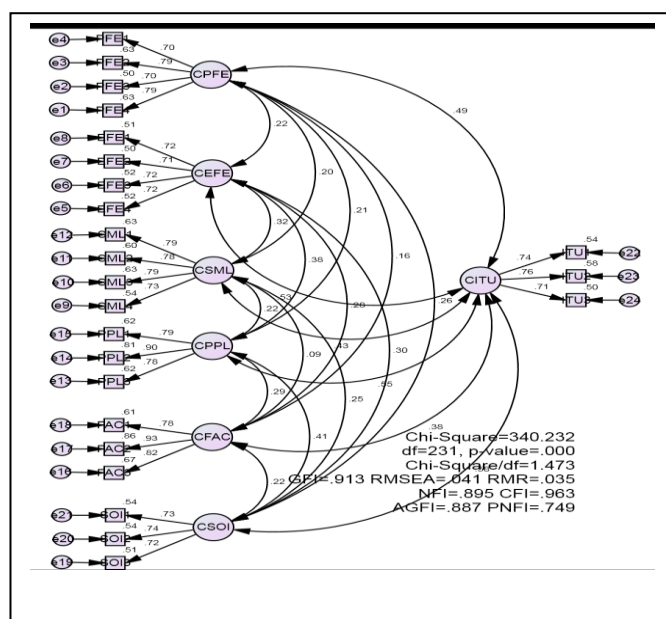


Fig. 2. Output of the measurement model

TABLE III. RESULTS OF CONVERGENT VALIDITY ASSESSMENT

Model construct	Item	Loading	Composite Reliability (CR)	Average Variance Extracted (AVE)
Intention to adopt	ITU1	0.737	0.780	0.542
	ITU2	0.761		
	ITU3	0.710		
Self Management Learning	SML1	0.734	0.857	0.600
	SML2	0.794		
	SML3	0.775		
	SML4	0.793		
Social Influence	SOI1	0.734	0.772	0.531
	SOI2	0.736		
	SOI3	0.715		
Facilitating Conditions	FAC1	0.780	0.881	0.713
	FAC2	0.928		
	FAC3	0.819		
Performance Expectancy	PEE1	0.704	0.837	0.563
	PEE2	0.792		
	PEE3	0.705		
	PEE4	0.794		
Perceived Playfulness	PPL1	0.789	0.865	0.682
	PPL2	0.898		
	PPL3	0.785		
Effort Expectancy	EFE1	0.717	0.808	0.512
	EFE2	0.707		
	EFE3	0.718		
	EFE4	0.721		

CR = (square of the summation of the factor loadings)/(square of the summation of the factor loadings) + square of the summation of the error variances); AVE = (summation of the square of the factor loadings) / (summation of the square of the factor loadings) + (summation of the error variances)

TABLE IV. RESULTS OF DISCRIMINANT VALIDITY ASSESSMENT

	[1]	[2]	[3]	[4]	[5]	[6]	[7]
[1] Intention to adopt	0.736						
[2] Self Management Learning	0.429	0.774					
[3] Social Influence	0.581	0.255	0.728				
[4] Facilitating Conditions	0.381	0.089	0.381	0.844			
[5] Performance Expectancy	0.486	0.202	0.486	0.202	0.750		
[6] Perceived Playfulness	0.550	0.225	0.550	0.225	0.550	0.825	
[7] Effort Expectancy	0.527	0.319	0.527	0.319	0.527	0.319	0.715

In Structural Equation Modelling, fit criteria are assessed in terms of absolute fit measures, incremental fit measures and also parsimony fit measures. As illustrated in Table 5, the χ^2 statistic suggests that the data do not fit the model well ($\chi^2 = 340.232$, $df = 231$, $p\text{-value} < 0.05$). However, because χ^2 is easily affected by sample size [15], the χ^2 statistic is not always an appropriate measure of a model's goodness-of-fit.

Therefore other fit indices as shown in Table V are used to examine the model's goodness-of-fit. Apparently, all fit indices surpassed the fit criteria suggesting that the SEM model fits the data very well.

TABLE V. FIT ASSESSMENT RESULTS

Fit Index	Fit Criteria	Measurement Model
Chi Square (χ^2)		340.232
Degrees of freedom		231
P-value (probability)	≥ 0.5	0.000
<i>Absolute fit measures</i>		
CMIN (χ^2)/DF	3	1.473
GFI (Goodness of Fit Index)	≥ 0.9	0.913
RMSEA (Root Mean Square Error of Approximation)	≤ 0.05	0.041
RMR (Root Mean Square Residual)	≤ 0.05	0.035
<i>Incremental fit measures</i>		
NFI (Normed Fit Index)	≥ 0.9	0.900
CFI (Comparative Fit Index)	≥ 0.9	0.963
<i>Parsimony Fit Measures</i>		
AGFI (Adjusted Goodness of Fit Index)	≥ 0.8	0.887
PNFI (Parsimonious Normed Fit Index)	≥ 0.5	0.749

Figure 3 displays the AMOS output of the structural model while Table VI showcases the path coefficients between the independent variables and dependent variable. The Squared Multiple Correlation (R^2) value for the relationship between the six independent variables and intention to adopt was 0.650. The overall results indicate that all hypotheses were fully supported as the p-values for all paths are well below 0.05. The coefficient values (β) range between 0.100 and 0.265. Figure 2 depicts the path diagram between the independent and dependent variables.

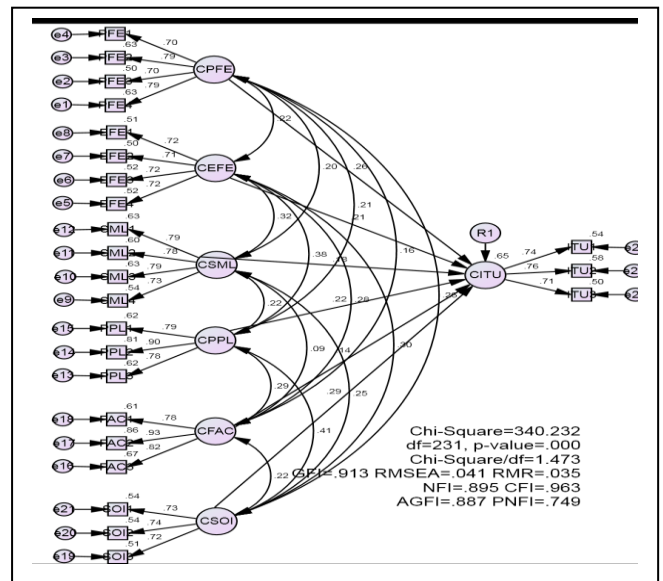


Fig. 3. Output of the structural model

TABLE VI. HYPOTHESES TESTING

Hypothesis	Coefficients	t-value	p-values	Supported
H1: Performance expectancy → intention to adopt	0.197	4.297	< 0.01	Yes
H2: Effort expectancy → intention to adopt	0.164	2.903	< 0.01	Yes
H3: Self management of learning → intention to adopt	0.146	2.909	< 0.01	Yes
H4: Perceived playfulness → satisfaction to adopt	0.184	2.448	< 0.01	Yes
H5: Facilitating condition → satisfaction to adopt	0.100	2.448	< 0.01	Yes
H6: Social influence → satisfaction to adopt	0.265	3.961	< 0.01	Yes

VI. DISCUSSION

The present study provides both a theoretical and practical contributions to understanding the predictors of intention to adopt mobile learning. The findings of this study should be of interest to both researchers and practitioners. The results generated from the path analysis indicate that the combination of the six independent variables accounts for 65% of the variance in intention to adopt mobile learning. This result suggest that 65% of the variance in intention to adopt mobile learning can be explained by performance expectancy, effort expectancy, social factors, facilitating conditions, perceived playfulness and self management of learning.

This study has significantly recognized the influence of performance expectancy on intention to adopt mobile learning ($\beta = 0.197, p < 0.01$). The result is consistent with [3], [10], [17], [22], [28], and [41]. The results suggest that, the more students perceive that mobile learning is useful for learning and improves their productivity; the more likely they are to engage in mobile learning. Theoretically, this result further strengthens UTAUT in predicting mobile learning adoption. The scale used for measuring performance expectancy focused on increased performance, productivity, and effectiveness. From the practical viewpoint, the findings send a strong message on the importance for increasing student performance expectancy. Educators and administrators could perhaps play a role by promoting the benefits and usefulness of mobile learning to their students and encourage them to use their mobile devices for information searching, engaging in online group discussions or completing other learning activities.

Just as performance expectancy, effort expectancy which is derived from UTAUT was also found to be a significant predictor of mobile learning adoption ($\beta = 0.164, p < 0.01$). The result is in tandem with that of [3], [10], [17], [22], [28], and [41], which means that the more students perceive that mobile learning is easy to use for learning; the more likely they are to engage in mobile learning. Effort expectancy construct is similar with perceived ease of use which is defined as the degree to which a person believes that the use of a particular system would be free of effort [8]. The items used for measuring effort expectancy focused on the degree of difficulty on using mobile learning. Today, among students of Malaysian universities, the use of mobile devices especially smart phones is very common. Perhaps, due to the fact that using a mobile device appears to be routine for most of these students; therefore they may perceive using it will not require much of their effort, as it is just similar to using it for other tasks. Nevertheless, this finding has provided additional support for UTAUT in predicting mobile learning. The implication to practitioner is that, when developing mobile learning applications, serious attention should be given on user-friendliness aspects.

The third hypothesis in this study is between self management of learning and intention to adopt mobile learning. Compared to the constructs of UTAUT, this variable is not very extensively studied in the context of mobile learning. The result of this study has showed that this construct is indeed applicable in determining intention to adopt mobile learning ($\beta = 0.146, p < 0.01$). This result is in line with the finding of [41]. This finding implies that individual with a highly autonomous learning ability will be more likely to use mobile learning than will an individual with a lower autonomous learning ability [41]. Given this finding, mobile learning developers should respond by developing mobile learning applications that are equipped with features that are suitable for those who are highly independent in their learning processes. On the other hand, educators and administrator can also play a role by grooming their students to be more independent and adapt themselves to be more self learning.

The results of this study also recognized perceived playfulness as a significant predictor of intention to adopt mobile learning ($\beta = 0.184, p < 0.01$). This finding further supports previous studies done by [10] and [41]. The result implies that the more students enjoy the mobile learning, the more they will be motivated to engage in mobile learning activities. [41] stated that, given that the usage of mobile learning is fully voluntary and that the target user group consists of a large number of people with very diversified backgrounds, making mobile learning system playful and enjoyable to interact with, is crucial for attracting more users to the mobile learning system. Therefore, mobile learning developers should react to this finding by enriching their mobile learning applications with enjoyable and entertaining features.

Consistent with [10], [17], and [28], this study has also found that facilitating condition as a essential predictor of intention to adopt mobile learning ($\beta = 0.184$, $p < 0.01$). This finding suggests that student will not be attracted to adopt mobile learning in the absence of facilitating conditions. In the context of Malaysia, all university students are entitled to a special voucher for purchasing smart phones. On top of that, the free wireless networks, available in the universities as well as in other public places such as bistros, restaurants and public libraries provide convenient internet access to the students. Nonetheless, this finding should alert the authorities concerned on the importance of continuous update and upgrade of the infrastructure or facilities required for the implementation of mobile learning.

The last construct being studied is social influence, which is also drawn from UTAUT. The results confirmed that social influence is a significant predictor of intention to adopt mobile learning ($\beta = 0.265$, $p < 0.01$). In fact, in this study, social influence is found to be the strongest predictors compared to other constructs. This result is also consistent with that of [3], [10], [17], [28], and [41]. Based on the result it can be concluded that the more students perceive faculty, peers, and other individuals important to them believe they should use mobile learning, the more likely they are to engage in mobile learning. Given this finding, it is crucial that people who have strong connection with the students such as the lecturers, colleagues or even family members, should persistently encourage the student to engage in mobile learning.

VII. CONCLUSION

The purpose of this article has been to explore factors that influence the intention of users to adopt mobile learning. To achieve this purpose, an empirical based framework drawn from UTAUT and previous empirical studies has been developed. The results of the analyses of the collected data significantly verified the established hypotheses. The results suggest that performance expectancy, effort expectancy, social factors, facilitating conditions, perceived playfulness and self management of learning are strong determinants of intention to adopt mobile learning.

Just like in any other research, there are several limitations associated to the conduct of this study. Firstly, is the choice of students that was confined to one university only. Future study, should consider extending the scope of population by taking students of other universities. Secondly, the study did not assess non-response bias on the questionnaire. In the future, researchers should also assess non response bias so as to increase the reliability and validity of the research findings.

ACKNOWLEDGMENT

The researcher would like to extend his thanks to the students from the Faculty of Information Management, Universiti Teknologi MARA who had willingly participated in this study. Special thanks are also owed to Accounting

Research Institute AND Universiti Teknologi MARA, Shah Alam Malaysia for providing all the necessary resources required for the completion of the study.

REFERENCES

- [1] A. Bandura, *Social Foundations of Thought and Action: A Social Cognitive Theory*. Prentice-Hall, NJ: Englewood Cliffs, 1986.
- [2] A. Murphy, *Mobile Learning in a Global Context: A Training Analysis*. Proceeding of the International Conference on Networking, International Conference on Systems and International Conference on Mobile Communications and Learning Technologies, 2006.
- [3] A.B. Nassuora, "Students Acceptance of Mobile Learning for Higher Education in Saudi Arabia," *International Journal of Learning Management Systems*, vol. 1, no. 1, pp. 1-9, 2013.
- [4] C. Fornell and D.F. Larcker, "Evaluating Structural Equation Models with Unobservable Variables and Measurement Error," *Journal of Marketing Research*, vol. 19, pp. 39- 50, 1981.
- [5] C.P. Schofield, T. West, and E. Taylor. *Going Mobile In Executive Education How Mobile Technologies Are Changing The Executive Learning Landscape*. Berkhamsted Hertfordshire Ashridge, 2011.
- [6] E.M. Rogers, *Diffusion of Innovations*. New York: Free Press, 1995.
- [7] F. Pozzi, *The Impact of M-Learning in School Contexts: An Inclusive Perspective*, em Stephanidis, C. (Ed.), *Universal Access in HCI, HCII, LNCS 4556*, Springer-Verlag, Berlin, 2007.
- [8] F.D. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," *MIS Quarterly*, vol. 13, no. 3, pp. 319-340, 1989.
- [9] F.M.S.M. Momani, and A.M. Abualkishik, "Factors Influencing Students' Intention to Adopt Mobile Blackboard," *International Journal of Science and Research*, vol. 3, no. 5, pp. 29-32, 2014.
- [10] F.Z. Hadi and A.A. Kishik, "Acceptance of Mobile Learning Among University Students in Malaysia," *Journal of Computing & Organizational Dynamics*, vol. 1, no. 1, 2014.
- [11] H. De Silva, D. Ratnadiwakara, and A. Zainudeen, "Social Influence in Mobile Phone Adoption: Evidence from the Bottom of the Pyramid in Emerging Asia," *Mobile Telephony Special Issue*, vol. 7, no. 3, pp. 1-18, 2011.
- [12] H.C. Kelman, "Compliance, identification, and internalization: three processes of attitude change," *The Journal of Conflict Resolution*, vol. 2, no. 1, pp. 51-60, 1958.
- [13] I. Ajzen, "The theory of planned behavior," *Organizational Behavior and Human Decision Processes*, vol. 50, pp. 179-211, 1991.
- [14] J. Chung and B. Tan, "Antecedents of Perceived Playfulness: An Exploratory Study on User Satisfaction of General Information Searching Websites," *Information & Management*, vol. 41, no. 7, pp. 869-881, 2004.
- [15] J.C. Anderson, and D.W. Gerbing, "Structural Equation Modelling in Practice: A Review and Recommended Two-Step Approach," *Psychological Bulletin*, vol. 103, no. 3, pp. 411-423, 1988.
- [16] J.F. Hair, B. Black, B. Babin, R.E. Anderson, and R.L Tatham, R.L. *Multivariate Data Analysis: A Global Perspective*. New Jersey: Pearson Education Inc., 2010.
- [17] K. Jairak, P. Praneetpolgrang and K. Mekhabunchakij, "An Acceptance of Mobile Learning for Higher Education Students in Thailand", *Proceedings of The Sixth International Conference on eLearning for Knowledge-Based Society*, 17-18 December 2009, Thailand.
- [18] K.A. Bollen, *Structural Equations With Latent Variables*. New York: Wiley, 1989.
- [19] L. Naismith, P. Lonsdale, G. Vavoula, and M. Sharples. *Mobile Technologies and Learning*. In *Futurelab Literature Review Series*, Report No 11, 2004.

- [20] M. Fishbein, and I. Ajzen. *Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research*. Addison-Wesley, MA: Boston Press, 1975.
- [21] M. Igbaria, S. Parasuraman, and J.J Baroudi. A Motivational Model of Microcomputer Usage. *Journal of Management Information Systems*, vol. 13, no. 1, 1996.
- [22] M. Jambulingam, "Behavioural Intention to Adopt Mobile Technology among Tertiary Students," *World Applied Sciences Journal*, vol. 22 no. 9, pp. 1262-1271, 2013.
- [23] M. Masrom, and R. Hussein, *User Acceptance of Information Technology: Understanding Theories and Model*. Selangor, Malaysia: Venton Publishing(M) Sdn. Bhd., 2008.
- [24] M.L.A. De Carvalho, H.C. de Azevedo Guimarães, and A.M.C. Gobbo, *Intention to Use M-Learning in Higher Education Settings*. Proceedings of the 36th Encontro De ANPAD, 22-26 September, 2012, Rio De Janeiro Brazil.
- [25] P.B. Muiyinda, "MLearning: Pedagogical, Technical and Organizational Hypes and Realities," *Campus-Wide Information Systems*, vol. 24, no. 2, pp. 97-104, 2007.
- [26] R. Agarwal and E. Karahanna, "Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage," *MIS Quarterly*, vol. 24, no. 4, pp. 665-694, 2000.
- [27] R.B. Kline, *Principles and Practice of Structural Equation Modeling*, 2nd ed. New York: Guilford Press, 2005.
- [28] S. Iqbal, and I.A. Qureshi, "M-learning Adoption: A Perspective From a Developing Country," *The International Review of Research in Open and Distance Learning*, vol. 13, no. 3, pp. 147-164, 2012.
- [29] S.J. Geddes, "Mobile learning in the 21st century: benefits for learners. Knowledge Tree e-journal," *An Ejournal of Flexible Learning in VET*, vol. 30, no. 3, pp. 1-13, 2004.
- [30] S.K. Sharma and F.L. Kitchens. *Web Services Architecture for M-learning*. *Electronic Journal for E-Learning*, vol. 2, no. 1, pp. 203-216, 2004.
- [31] T. Evans, *Flexible Delivery And Flexible Learning: Developing Flexible Learners?* In V. Jakupec and J. Garrick (Eds), *Flexible learning, human resource and organizational development*. London: Routledge, 2000.
- [32] Y. Liu, *Solving the Puzzle of Mobile Learning Adoption*. Unpublished Doctoral Thesis, Department of Information Technologies, Faculty of Technology, Åbo Akademi University, 2011 Available at: https://www.doria.fi/bitstream/handle/10024/69575/liu_yong.pdf?sequence=1
- [33] Y.S. Poong, A. Yamaguchi, and J.I. Takada, *Determinants of Mobile Learning Acceptance in Luang Prabang: Towards World Heritage Site Preservation Awareness Promotion*. Proceedings of the 24th International CIPA Symposium, 2 – 6 September 2013, Strasbourg, France.
- [34] P.J. Smith, K.L. Murphy, et. al. "Towards Identifying Factors Underlying Readiness for Online Learning: An Exploratory Study," *Distance Education*, vol. 24, no. 1, pp. 57-67, 2003.
- [35] B. Suh, and I. Han, "Effect of trust on customer acceptance of Internet banking," *Electronic Commerce Research and Applications*, vol. 1, pp. 247-263, 2002.
- [36] J.P. Tan, and E. McWilliam, "Cognitive Playfulness, Creative Capacity and Generation of CW Learners," *Cultural Science*, vol. 1, no. 2, pp. 1-7, 2008.
- [37] S. Taylor and P.A. Todd, "Assessing IT Usage: the role of prior experience," *MIS Quarterly*, vol. 19, no. 2, pp. 561-570, 1995.
- [38] R.L. Thompson, C.A. Higgins, and J.M. Howell, "Personal computing: Toward a conceptual model of utilization," *MIS Quarterly*, vol. 15, no. 1, pp. 124-143, 1991.
- [39] V. Venkatesh, "Determinants of Perceived Ease of Use: Integrating Control, Intrinsic Motivation, and Emotion Into the Technology Acceptance Model," *Information Systems Research*, vol. 11, no. 43, 342-365, 2000.
- [40] V. Venkatesh, M.G. Morris, G.B. Davis, and F.D. Davis, "User Acceptance of Information Technology: Toward a unified view," *MIS Quarterly*, vol. 27, no. 3, pp. 425-478, 2003.
- [41] Y.S. Wang, M.C. Wu, and H.Y. Wang, "Investigating The Determinants and Age and Gender Differences in the Acceptance of Mobile Learning," *British Journal of Educational Technology*, vol. 40, no. 1, pp. 92-118, 2009.
- [42] D. Warner, G. Christie, and S. Choy, *The Readiness of the VET Sector for Flexible Delivery Including On-Line Learning*. Brisbane: Australian National Training Authority, 1998.

Predictors of E-Participation Levels: The Case of Jordan

Heba K. Al-Quraan

MIS dept., IT College, Yarmouk University
Irbid, Jordan

heba.quraan@yahoo.com

Emad A. Abu-Shanab

MIS dept., IT College, Yarmouk University
Irbid, Jordan

abushanab@yu.edu.jo

Abstract: According to the revolutionary advancement in information and communication technologies, citizens are becoming more open-minded, ambitious, aware of technology capabilities, and empowered enough to participate in the decision making process. Citizens are motivated to be an active part in the political process and they are encouraged to be involved in order to have collected feedback from them for the social and political reform process. This study tried to explore the factors influencing the level of electronic participation (e-participation) within an e-government context. Ease of use, infrastructure readiness, cost, and relative advantage are factors hypothesized to have an impact on participation success. The results supported the influence of three predictors (ease of use, cost, and relative advantage) in their effect on participation levels. Only infrastructure readiness was not significant in predicting participation levels. Conclusion and future work are discussed later in the paper.

Keywords: E-government, Open government, E-participation, E-consulting, Social Media, Transparency, Collaboration, ICT, Digital Divide, Maturity Model.

1. INTRODUCTION

E-government is an extension of traditional government that supports conducting all related transactions electronically. E-government implementation enhances performance, reduces cost, minimizes errors, utilizes information and communication technology (ICT), improves service provision process, and increases transparency and credibility [1]. E-government also empowers citizens to participate in the decision making process and policy making to guarantee the highest participation and interaction from citizens. E-government includes the following dimensions: e-democracy, e-consultation, e-campaign, e-voting, e-election, and e-participation. E-participation means using any electronic channel/device to contribute in the political process. Research reported more than one version of e-participation levels, where three, four and five stages were reported from different perspectives. The major five stages are: 1) *E-informing*: it is about providing and publishing information to the public by the government in a one way interaction mode. 2) *E-consulting level* is the second one, which represents a two way interaction between citizens and governmental bodies in order to take feedback from them. 3) *E-involving* is the third level that aims at keeping citizens in touch (involved) with every decision made and to understand their concerns and issues. 4)

E-collaborating is the fourth level, which represents more advanced two way interaction and revolves around the partnership between governments and the public. 5) *E-empowering* level is the level through which citizens reach the highest engagement in the political process and take the responsibility of taking their own decisions, it is the last level [2].

The paper explored the levels of e-participation, and the literature related to such concept. Also, an empirical test was conducted utilizing a sample of responses to probe Jordanians' perceptions towards such phenomenon and the factors influencing it. The paper consists of a literature review (section 2) followed by a description of the research method conducted. Section 4 will describe the analysis and discussion of results. Finally, conclusion and future work are depicted at the end.

2. LITERATURE REVIEW

One of the major objectives of the e - government projects is to reach the largest number of audience to be involved in policy and decision making processes. Governments need to collect public opinions, make them satisfied with such initiative; and to assure the project acceptance and its success [3]. Furthermore, governments are utilizing social media for

reaching citizens and enhance the participation level. Many researchers consider such initiative as a mean to achieve mutual benefit between the government and citizens; for the government, it improves efficiency, effectiveness, and facilitates service delivery process. For citizens, e-government plays significant role in increasing transparency, accessibility, accountability, and their participation in the political process [4].

2.1 E-government and e-participation definitions

E-government is defined as the existence of required and special services and options of websites, which focus on online service delivery [5]. Another definition focus on utilizing ICT tools in order to facilitate the interaction between citizens, businesses, and the other governmental agencies [6] [7].

E-government is considered as a national project that has been developed by the aid of both public and private sectors; which have the responsibility to promote more efficient and effective government. It also facilitates more accessible services, and makes government more accountable to citizens [8]. Utilizing an empirical test, Abu-Shanab has emphasized the importance of the three major dimensions related to the success of e-government projects and they are summed in the following: infrastructure readiness, social forces and governmental issues [9].

Anthopoulos, Siozos, and Tsoukalas have illustrated that e-government initiative is a project that may include policies and targets, but not principles or even instructions. Also, such project needs to be designed and then executed carefully by involving civil servants through a bottom-up design process, which aims at enhancing the participation and knowledge sharing processes by using e-Government Groupware (eGG) application (a collaboration application that exploits the digital public services) [10].

The authors have defined e-government project as " The use of ICT in an innovative way; to exploit any opportunity that leads to enhance the relationship between government and citizens and making them empowered enough to participate effectively in the political process in its different forms and patterns, and enabling government to have constructive feedback from citizens to participate efficiently in both social as well as political reforming processes."

E-participation is considered as a dimension or a subset of e-government as a whole; it is seen as the engagement and involvement of citizens in the decision making process, by benefitting from the advantages of ICT tools in order to improve social and political responsibility [11]. E-government and e-participation are related within two aspects: the first association is related to the political system and the society interaction, where e-participation represents public policy formulation process. On the other hand, the relationship between society and the administrative system can be considered as e-government initiative itself, where it

emphasizes the improvement of service provision process [12].

2.2 Case Studies in e-participation

The Arab Open Government is a study that focuses on appreciating the role of technology represented by social media specially the distribution of information to citizens so they can be an effective part in the open government initiative. The initiative emphasizes using one way interaction tools to achieve its objectives in reflecting transparency, public participation, service provision process, and collaboration. The study also indicated that there is an initial interest of open government project in Bahrain, Saudi Arabia, UAI, Morocco, and Jordan [13].

The United Arab Emirates (UAE) case study has explored the readiness of infrastructure to assure the success of the e-government project. It has clarified a number of stages and one of them was the interaction and participation. The first is *emerging*, which is about having an official website for the government. The second is about how the stored political information can be accessed by the citizens and called it *enhanced*. The third one is the core of this paper which revolves around delivering convenient services to citizens via a friendly website or portal; it was called *interactive*. The *transactional* stage is the fourth one which is about introducing two way interaction channel to connect the governments with citizens. The final stage is *connected*, which prepares the back office infrastructure to be able to respond quickly to citizens' needs and expectations [7].

In the UK, e-government is considered as an information system (IS) that has been designed to exploit any opportunity to enhance the communication of citizens in the democratic process, thus improve the level of e- participation [14].

Based on the previous cases, the authors suggest that the relationship between the interactivity and e-participation is much supported. Once citizens are informed by the government with any needed political information (like providing their sensitive data, suggestions, or even their complements/complaints to the government), both parties are getting closer to each other and thus eliminate the gap between them. Citizens' interactions and engagements with the e-government program is the best form of participation, which supports a positive correlation between interactivity and e-participation.

2.3 E-participation evaluation models

The UN E-Participation Index was developed as a qualitative indicator of both the capacity and the willingness in encouraging citizens in promoting and enhancing a participatory decision-making process and to reach out for the citizens within its own governance program. E-participation index attempts to see whether the country is able to increase e-information for citizens, promote e-decision making, and enhance the e - consultation process. Curtin attempted to design an e-participation model that focused on providing citizens with enough information to participate efficiently and effectively in policy and decision making process. The

proposition tries to enable them to express their opinions and suggestions to be heard and known [15].

The research proposed five key indicators of e-government projects and they are: service, technology, employees, policy and social responsibility (STEPS model). In their paper, the research model included a number of these factors or sub factors; (i.e. Infrastructure is included as a sub-dimension within the technology factor; and the cost is a sub-factor of policy). Relative advantage is embedded in the service dimension and includes support and efficiency. Finally the participation factor was estimated by user take-up factor. The result of the study supported the influence of the five predictors [16].

An analytical framework of ten dimensions was proposed and studied to determine citizens' participation, technologies, and the level of participation, which represents how far citizens can be engaged in the democratic process. The study also explored the decision making process and the rules of engagement and needed information. The other five dimensions are duration, accessibility, resources, evaluation, and critical success factors [17].

Navarra and Cornford looked at how e-government services, applications, and infrastructure can be developed. They proposed to study multiple models from four dimensions (governance model, service delivery, policy focus, and the actors) which are managerial, consultative, disciplinary, and the participatory models. The last model represents the promotion of free speech and expression. It revolves around the civil society involvement executed and supported by the voluntary associations [18].

2.4 Innovations in ICT to Promote E-Participation

ICT plays significant role in achieving the transparency and credibility, and it is an effective mean to increase the openness and to reduce the corruption by enabling citizens to track their own activities and decisions instead of supporting them with final information. A number of key factors for building a culture of transparency to increase citizen participation in the e - government initiative and overcome any potential barrier through the combination between technology and the political will; these factors are ICT access, trust, empowerment, social capital, and the acceptance of transparency itself [19].

For the importance of e - participation issues, many researchers were interested in deciding which technologies are the most appropriate to meet the goals and objectives of involving citizens in the policy making in the democratic process, Macintosh had explored and clarified ten key dimensions (which have been mentioned in the previous section) for characterizing e-participation in policy making and focused on utilizing ICT in an innovative way to get the intended aims [17]. Governments should take some pro-active measures to guarantee reaching the largest audience base, and make ICT-related services accessible, attainable, and available anytime. Cell phones, PDAs, wireless networking, and speech technology are efficient tools to promote e-participation. There are three major prerequisites for e-participation: governments

should focus on targeting specific issues, specific groups, and selecting a small number of priorities [20].

Phang and Kankanhalli proposed a framework that focuses on studying ICT utilization toward e-participation initiatives. They concluded that not a single participatory technique or ICT tool can satisfy multiple goals, where each objective has its own means to be achieved effectively. Countries may force e-government projects to be a path for openness, but such decision needs to be taken cautiously; as the shift from traditional government toward e-government requires tailored user involvement practices in government development projects [21].

2.5 E-Participation and the digital divide

The "digital divide" or "digital inequality" phenomenon is attracting more attention from researchers as it is the path to the success of e-government projects. The digital divide has five major types: demographic divide, which may include gender, age, education, location of resident factors. The second type is the economic divide, which is related to socioeconomic circumstances. The access divide is the third type that is concerned with the physical access to ICT. The fourth one is the capability divide, which focuses on the citizen's believe of his or her capability to utilize and benefit from ICT. The last type is the innovativeness divide, which relates to the tendency to change and try new/different technologies [22].

Al-Rababah and Abu-Shanab have studied the demographic divide and focused on exploring the impact of gender digital divide and the participation of women in the e-government domain. The study concluded that little or even no IT skills are acquired by women in rural areas, and thus need to empower them to gain access to e-government services [23]. They concluded that there is still a gap between actual participation in e-government systems and the required and intended level of participation. Usability and accessibility are key factors that come to reduce and simplify the digital divide and they could be related with creating and increasing the divide itself (including age, gender, culture, income, education, and disability) [24].

In the case of United States (US), many strategies were developed in order to secure Internet access, implement the needed training and education, and encourage using the Internet to increase the participation in the e - government project. Such efforts are targeted toward disadvantaged areas as a proposed solution for bridging the digital divide problem [25]. The digital divide can be considered as one of the reasons for increasing the failure rate of e-government projects. This virtual inequality is a real limitation that hinders such projects. Many potential procedures can be put in place to attempt to eliminate such factor by having a comprehensive view of the policies proposed, and taking the issues of supply and demand into account. Such procedure may have a positive impact on e-government within societies [26].

The digital divide is a challenge for e-government project adoption, where research has clarified four key barriers to the

access and use of e-government systems, they are the awareness, trust, usefulness, and digital divide [27] [28]. Rahman has studied the role of trust and the digital divide within the context of e-government and concluded with the following: trust is an important factor in reducing the complexity by converging the different expectations. Trust also is a factor believed to be an indicator for e-government systems success. The author asserted the need to develop new comprehensive policy to bridge the gap and improve e-government success [29]. Another study tried to explore the relationship between the quality of e-government services which has been reflected by (communication, collaboration, openness, and sharing of information factors) and the trust in it according to the digital divide groups (passive, progress, alienation, and desire groups of information). The result was that there is a partial correlation between quality and trust of e-government services. Also, there should be policies to be developed to overcome the obstacles related to information sharing factor, and keeping control on issues like; privacy, security, and the collaboration among individuals [30].

3. RESEARCH METHODOLOGY

The literature review concluded with a set of factors that would influence the e-participation levels. Four key factors would increase the levels of participation electronically within e-government context. Such factors would enable citizens to become active component of such system and being empowered enough for interactive participation. Table 1 depicts the definition of variables proposed in this study, and Figure 1 illustrates the research model proposed based on the literature review and the previous studies. Based on that, the following hypotheses are stated:

H1: Ease of use of e-government systems will be positively associated with the expected e-participation levels.

H2: Relative advantage of e-government systems will be positively associated with the expected e-participation levels.

H3: The cost of e-government systems will be positively associated with the expected e-participation levels.

H4: Infrastructure readiness will be positively associated with the expected e-participation levels.

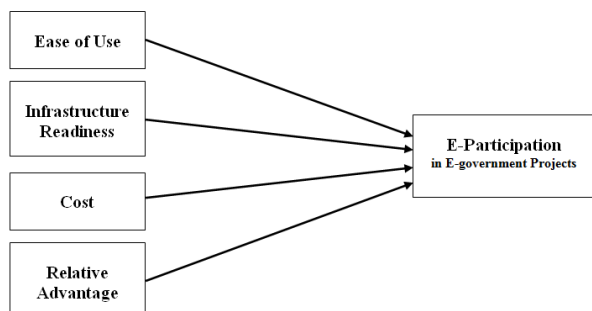


Figure 1: proposed research model

Table 1: Proposed definitions of Constructs

Construct	Definition
Ease of use	The use of friendly technologies, and the utilization of other facilities in an easy and compatible fashion.
Readiness	The level of availability of needed infrastructure components (e-devices, PCs, and network connections.)
Cost	The reduced prices of the devices needed to use e-government services.
Relative advantage	The expected generated benefits to be achieved from using such services by citizens.
E-participation	The involvement of citizens in the political process (policy and decision making processes) and their role in the development process by their interaction.

3.1 Instrument used

The paper used previous research to build the instrument used in this study. A questionnaire included three sections was utilized to collect the required data of this research. The first section clarified the purpose of research and the concepts of e-government and e-participation. The second section included few demographic questions like gender, age and education. The instrument included no identification questions exposing respondents identity and to let respondents freely fill the survey. The final section included 22 items measuring the five constructs shown in the research model. The five constructs used in this study were measured using: 4 items for measuring ease of use, 4 items for measuring infrastructure, 4 items for measuring cost, 4 items for measuring relative advantage, and 6 items for measuring participation.

The instrument utilized a 5 point Likert scale with 1 representing total disagreement, and 5 representing a total agreement. Once the instrument was ready, content analysis was conducted using 5 master degree students who are taking a course in e-government. They read the survey and commented on both the statements and the content, then they added their comments on the items and the dimension they are measuring.

The instrument was also checked for internal reliability and yielded acceptable results. Cronbach's alpha was used for measuring internal consistency, where all values were more than 0.6 (the threshold level for such measure in social sciences.) Table 2 shows the estimations of Cronbach's alpha for the five variables in the model. The highest value was for the participation construct, and the lowest value was to relative advantage, but still at acceptable levels.

Table 2: Reliability Analysis and the values of Cronbach's alpha

Construct	Valid N	Number of	Cronbach's

		Items	Alpha
Ease of use	248	4	0.842
Infrastructure	236	4	0.863
Cost	245	4	0.778
Relative advantage	239	4	0.775
Participation	237	6	0.903

3.2 Sample and sampling process

The survey was distributed to many schools in the Northern part of Jordan, where teachers filled the survey and return it to the researchers by hand. Teachers were chosen as a sample because they represent the largest and the most educational segment using government services. Such sample is also convenient to our study from two sides: first, it is easy to attain the required sample size. Second, their sociability and their ability to express their perspectives within the domain they are in. Also, building an online survey contributes in widening the size of the sample, around 100 surveys were received online via email and google drive. The survey collection process was done in August 2014, and the total surveys distributed were 260 surveys. The total collected surveys were 250 surveys, were 2 surveys were removed because of empty responses (only demographic data filled). The total size of used sample was 248 surveys, in which the male participants were with 51.6% and female with 48.4%. The majority of them was between 20 and 40 years old with 69.8% and 58.5% of them are having Bachelor degree. Table 3 shows the demographic of sample used.

3.3 Descriptive statistics of variables

The first step in the analysis was to investigate the level of perceptions Jordanians hold for the five variables in the instrument. Table 4 shows the results of calculating the means and standard deviations for the five variables in the research model. Results indicated high perceptions for the five variables as all means were above 3.66. The cutoff points for a 5 point Likert scale in the social sciences is 1-2.33 representing low perceptions, 2.33-3.66 as medium perceptions, and 3.66-5 as high perceptions. The results indicated that all constructs are highly perceived and thus support our research premise and the importance of the constructs adopted. Finally, the standard deviation values indicate a consensus with close values when comparing the variables.

Table 3: Sample demographics

Gender	Count	%
Male	128	51.6%
Female	120	48.4%
Total	248	100%
Age	Count	%
Less than 20 years	22	8.9%
20-40 years	173	69.8%
41-60 years	51	20.6%
More than 60 years	1	0.4%
Not reported	1	0.4%
Total	248	100%
Education	Count	%

High School or less	44	17.7%
Bachelor	145	58.5%
Master/PhD	49	19.8%
Other	10	4.0%
Total	248	100%

Table 4: The means and standard deviations for the research model variables

Construct	N	Min	Max	Mean	Std. Deviation
Ease of use	248	1	5	4.002	0.734
Infrastructure	236	1	5	3.781	0.808
Cost	245	1	5	3.952	0.825
Relative Advantage	239	1	5	3.803	0.757
Participation	237	1	5	3.925	0.780

4 DATA ANALYSIS AND DISCUSSION

The proposed research model hypothesized four relationships that map the factors influencing the levels of electronic participation in e-government projects. The data collected was analyzed using SPSS and the research model was tested using multiple regression, where four independent variables predicted one dependent variable. The results of the model indicated a high prediction value of e-participation, with a value of coefficient of determination $R^2 = 0.693$ (adjusted $R^2 = 0.688$). The model was extremely significant with a value of $F_{4,244} = 125.513$, $p < 0.001$. Such high value indicates a great explanation of the dependent variable equal to 69% of the variance in sustainability. The model also utilized four predictors that define the four stated hypotheses. To explore the individual influence of the predictors, a coefficient table was generated which is shown in Table 5 below.

The results in the table indicate that three of the four constructs significantly predict e-participation. The only construct that failed to predict the dependent variable is the infrastructure readiness. The recommendations of research in the area of e-readiness emphasized the role of infrastructure in improving e-government accomplishments in the country and specifically the relationship with businesses [31]. Still, the other three variables significantly predicted participation at the 0.05 level. Ease of use and cost factors predicted participation at the 0.05 level, whereas, relative advantage predicted participation at the 0.001 level.

Table 5: The coefficient table of the regression test

Construct	Unstand. Beta	Standard Error	Stand. Beta	t	Sig.
Ease of use	0.123	0.056	0.116	2.184	0.030
Infrastructure readiness	0.021	0.060	0.023	0.351	0.726
Cost	0.144	0.062	0.152	2.342	0.020
Relative advantage	0.621	0.068	0.608	9.197	0.000

Dependent Variable: e-Participation

5 CONCLUSIONS AND FUTURE WORK

E-participation is one of the major dimensions of open government initiative proclaimed by the Obama's administration [32]. This research tried to investigate Jordanians level of e-participation in e-government project as reported by a sample of 248 citizens. Jordanians perceived highly the five proposed variables utilized in this study. Also, the five variables demonstrated reliable estimates.

The major factors influencing the level of e-participation directly were regressed on the dependent variable (e-participation in e-government projects) and they are: ease of use of e-government systems, which plays a significant role in encouraging citizen to be motivated and involved in the political arena through the interactive participation in e-government services and websites. Secondly, the cost of using such e-service is expected to be lower than the cost of getting the same service traditionally. Thirdly, the expected advantage from using such systems is increased through the increased participation and by using e-government systems. All these factors were supported in the study, but the influence of infrastructure readiness was not supported. It is believed that there is no well equipped and prepared infrastructure to start launching the initiative of e-government (not ready yet).

Based on this, three out of four hypotheses were supported, where H2 was the only rejected hypothesis. More research is required to investigate the real reasons behind citizens' perception about the deficiency of the infrastructure in Jordan. Also, a validation of the instrument using a different sample will provide future research with a suitable instrument to be used for measuring the five variables used.

Biography

Heba K. Al-Quraan earned her both master and Bachelor degrees in MIS from Yarmouk University (YU) in Jordan. Her research interest in areas like *E-government*, *technology acceptance*, *E-learning*, and *medical informatics*.

Dr. Emad A. Abu-Shanab earned his PhD in business administration, in the MIS area from Southern Illinois University – Carbondale, USA, his MBA from Wilfrid Laurier University in Canada, and his Bachelor in civil engineering from Yarmouk University (YU) in Jordan. He is an associate professor in MIS. His research interest in areas like *E-government*, *technology acceptance*, *E-marketing*, *E-CRM*, *Digital divide*, and *E-learning*. Published many articles in journals and conferences, and authored three books in e-government. Dr. Abu-Shanab worked as an assistant dean for students' affairs, quality assurance officer in Oman, and the director of Faculty Development Center at YU.

REFERENCES

- [1] Abu-Shanab, E. (2014). *Electronic Government, a tool for good governance and better service*. Dar Al-Ketab, second edition (ISBN: 978-9957-550-99-8), 2014, 262 pages.
- [2] Al-Dalou', R. & Abu-Shanab, E. (2013). *E-Participation Levels and Technologies*. The 6th International Conference on Information Technology (ICIT 2013), 8-10 May, 2013, Amman, Jordan, pp.1-8.
- [3] Alcaide-Muñoz, L., Hernández, A. M. L., & Caba-Pérez, C. (2014). *Public Managers' Perceptions of e-Government Efficiency: A Case Study of Andalusian Municipalities*. In *Measuring E-government Efficiency*, pp. 135-156. Springer New York, pp.1-19 "No issue and volume determined"
- [4] Tufts, S. H. (2014). *Citizen Engagement 2 Facebook Pages To De Civic Engagment*. *Journal of Information Technology Management*, 25(2), 16, pp. 15-21.
- [5] Rorissa, A., Demissie, D., & Pardo, T. (2011). *Benchmarking e-government: A comparison of frameworks for computing e-government index and ranking*. *Government Information Quarterly*, 28(3), 354-362.
- [6] Joseph, R. C. (2013). *A structured analysis of e-government studies: Trends and opportunities*. *Government Information Quarterly*, 30(4), 435-440.
- [7] Westland, D., & Al-Khouri, A. M. (2010). *Supporting e-government progress in the United Arab Emirates*. *Journal of E-Government Studies and Best Practices*, 2010, pp. 1-9.
- [8] Odat, A. & Khazaaleh, M. (2012). *E-Government Challenges and Opportunities: A Case Study of Jordan*. *IJCSI International Journal of Computer Science Issues*, 9(5), 2, September 2012, PP. 361- 367
- [9] Abu-Shanab, E. (2012). *Digital Government Adoption in Jordan: An Environmental Model*. *The International Arab Journal of e-Technology (IAJeT)*, Vol. 2(3) January 2012, pp. 129-135.
- [10] Anthopoulos, L. G., Siozos, P., & Tsoukalas, I. A. (2007). *Applying participatory design and collaboration in digital public services for discovering and re-designing e-Government services*. *Government Information Quarterly*, 24(2), 353-376.
- [11] Maier-Rabler, U., & Huber, S. (2010). *Sustainable e-participation through participatory experiences in education*. *JeDEM-eJournal of eDemocracy and Open Government*, 2(2), 131-144.
- [12] Peristeras, V., Mentzas, G., Tarabanis, K. A., & Abecker, A. (2009). *Transforming E-government and E-participation through IT*. *Intelligent Systems, IEEE*, 24(5), 14-19
- [13] Schwalje, W., & Aradi, W. (2013). *An Arab open government maturity model for social media engagement*. Tahseen Consulting. Retrieved January, 19, 2013, pp. 1-35
- [14] Olphert, W., & Damodaran, L. (2007). *Citizen participation and engagement in the design of e-government services: The missing link in effective ICT design and delivery*. *Journal of the Association for Information Systems*, 8(9), 27, pp. 492-507
- [15] Curtin, G. G. (2006). *Issues and Challenges: Global EGovernment/E-Participation, Models, Measurement and Methodology*. *E-Participation and E-Government: Understanding the Present and Creating the Future*, pp. 1-33

- [16] Osman, I. H., Anouze, A. L., Azad, B., Daouk, L., Zablith, F., Hindi, N. M., & Weerakkody, V. (2013). The elicitation of key performance indicators of e-government providers: A bottom-up approach, pp. 1-17
- [17] Macintosh, A. (2004, January). Characterizing e-participation in policy-making. In System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on (pp. 1-10).
- [18] Navarra, D. D., & Cornford, T. (2005). ICT, Innovation and Public Management: Governance, Models and Alternatives for eGovernment Infrastructures, pp. 1-11
- [19] Bertot, J. C., Jaeger, P. T., & Grimes, J. M. (2010). Using ICTs to create a culture of transparency: E-government and social media as openness and anti-corruption tools for societies. *Government Information Quarterly*, 27(3), 264-271.
- [20] Ahmed, N. (2006, April). An overview of e-participation models. In UNDESA workshop “E-participation and E-government: Understanding the Present and Creating the Future”, Budapest, Hungary (pp. 1-17).
- [21] Phang, C. W., & Kankanhalli, A. (2008). A framework of ICT exploitation for e-participation initiatives. *Communications of the ACM*, 51(12), 128-132.
- [22] Rahman, A. (2014). Toward a comprehensive conceptualization of the digital divide and its impact On e-government system success: evidence from local governments in Indonesia, pp. 1-26
- [23] AL-Rababah, B. & Abu-Shanab, E. (2010). E-Government and Gender Digital Divide: The Case of Jordan, *International Journal of Electronic Business Management (IJEEM)*, V8(1), 2010, pp. 1-8.
- [24] Choudrie, J., Ghinea, G., & Songonuga, V. N. (2013). Silver surfers, e-government and the digital divide: An exploratory study of UK local authority websites and older citizens. *Interacting with Computers*, iws020, pp. 1-25
- [25] Sipior, J. & Ward, B. T. (2005). Bridging the Digital Divide for e-Government inclusion: A United States Case Study. *The Electronic Journal of e-Government*, 3(3), pp. 137-146, available online at www.ejeg.com
- [26] Helbig, N., Ramón Gil-García, J., & Ferro, E. (2009). Understanding the complexity of electronic government: Implications from the digital divide literature. *Government Information Quarterly*, 26(1), 89-97.
- [27] Abu-Shanab, E. & Al-Azzam, A. (2012). Trust Dimensions and the adoption of E-government in Jordan. *International Journal of Information Communication Technologies and Human Development*, Vol. 4(1), 2012, January-March, pp.39-51.
- [28] Abu-Shanab, E. & Abu-Baker, A. (2011). Evaluating Jordan’s E-government Website: A Case Study. *Electronic Government: An International Journal*, Vol. 8(4), 2011, pp. 271-289.
- [29] Rahman, A. (2014). Does Trust in E-Government Mediate the Relationship between Digital Divide and E-Government Use?. *Middle-East Journal of Scientific Research*, 21(8), 1203-1212.
- [30] Myeong, S., Kwon, Y., & Seo, H. (2014). Sustainable E-Governance: The Relationship among Trust, Digital Divide, and E-Government. *Sustainability*, 6(9), pp. 6049-6069.
- [31] Obeidat, R. & Abu-Shanab, E. (2010). Drivers of E-government and E-business in Jordan. *Journal of Emerging Technologies in Web Intelligence*, Vol. 2 (3), August 2010, pp. 204-211.
- [32] Abu-Shanab, E. (2015). Open Government Initiatives In Public Sector: A Proposed Framework For Future Research. *Saba Journal of Information Technology and Networking*, Vol. 3(1), pp. 4-14.

The Development of Software Agents in e-Learning 3.0

Dorota Jelonek

Faculty of Management
Czestochowa University of Technology
Czestochowa, Poland
jelonek@zim.pcz.pl

Abstract— E-learning has significantly changed the process of educating students and employee training. The purpose of this article is to analyze the development trends of software agents used in e-learning. Particular attention is given to selected properties of agents, which can improve the services offered by agents assisting student and agents supporting teachers. Furthermore, the article presents the advantages and disadvantages of e-learning and an evolution of e-learning from the perspective of: e-learning 1.0, e-learning 2.0 and e-learning 3.0.

Keywords — e-learning 1.0; e-learning 2.0; e-learning 3.0; intelligent agents

I. INTRODUCTION

E-learning has been widely used for university-based and enterprise-based education. It is gaining applicability as an educational tool for a cost savings, institution reusability, its ability to enable students to study without the constraints of time and space and learner flexibility. In the literature of subject there is still ongoing discussion over advantages, disadvantages, quality, improvement of technical solutions and efficiency of e-learning application. In the discussion participate such experts from different domains as: education, computer science, information systems, psychology, sociology and educational technology, due to the fact that only multidimensional perception of e-learning can ensure success of such ventures [1].

E-learning appeared at the beginning of 90-ties of XX century and is constantly developed both in traditional e-learning form, hybrid training which is blended learning [2], m-learning (mobile learning) [3], b-learning (bloglearning) [4] and g-learning (game learning) [5].

The development of internet technologies that are used in e-learning allows to describe it from several perspectives: e-learning 1.0, e-learning 2.0 and e-learning 3.0 [6], [7], [8].

For building the e-learning 3.0 systems can be used new technologies as: Big Data or global data repository, linked data, cloud computing, smart mobile technology, personal avatars, 3D visualization, Semantic web and artificial intelligence e.g. intelligent agent.

Intelligent agent can be used in e-learning applications in different contexts. The various agent properties like autonomy, mobility, proactive and reactive behaviors, capability to co-

operate and communicate with other agents makes it ideal for use in e-learning.

II. ADVANTAGES AND DISADVANTAGES OF E-LEARNING

The concept of e-learning functions in many contexts and includes a wide range of definitions. In the educational approach, e-learning is a way of teaching, education supported by digital technologies. This aspect is underlined also by the definition: “e-learning is the use of new multimedia technologies and the Internet to improve the quality of learning by facilitating access to resources and services, as well as remote exchange and collaboration” [9]. In many publications there are emphasized technical and technological conditions of e-learning. Such a view is presented by the definition: “e-learning is the use of electronic media for a variety of learning purposes that range from add-on functions in conventional classrooms to full substitution for the face-to-face meetings by online encounters” [10].

Characteristics of a distance learning system [11]:

- Individual learning
- Individual pace
- Arbitrary learning time
- Arbitrary learning place
- Student takes responsibility
- Interactivity
- Physical separation
- Illimitability of time and space
- Self-verification of knowledge
- Student cooperation
- One-to-one consulting.

Therefore, elementary characteristics of distance learning are its practicality, efficiency and flexibility.

Several mentioned above characteristics of e-learning could be perceived as advantages while the others as disadvantages. For comprehensive description of e-learning both advantages and disadvantages will be identified.

The advantages of e-learning can be considered as the follows [12]:

- Student can study anywhere as long as there is access to a computer with internet connection,
- They can work at own pace,
- User can accommodate different learning styles through different activities,
- Flexibility to join discussions any hour of the day,
- E-learning is cost effective.
- Convenience and flexibility.
- Reviewing material.
- Student motivation. Some students may find asynchronous online work more engaging, as they can interact with the material when they are freshest and most productive.
- Fewer pressures on limited space. Online education can reduce pressure on university facilities by freeing up classrooms.
- Analytics and assessment.
- Access and support. Online classes provide vital access to place-bound populations and other groups traditionally underserved by institutions of higher education,

Unlike print media, e-learning can also provide individualized instruction, and instructor-led courses allow clumsily and at great cost. In combination with evaluating needs, e-learning can target specific needs. By using learning style tests, e-learning can help to locate and target individual learning preferences. What is more, synchronous e-learning is self-paced.

The disadvantages of e-learning training are represented from different aspects [13] [12]:

- Lack of personal community and connection (not for blended learning),
- Its a banking model of education (which is partially inevitable),
- Not necessary based on the best science regarding How People Learn,
- Tech, toys, and teaching over learning,
- Focus on memorization over learning core competencies,
- Better aligning of incentives of teachers and learners,
- Downtime plus mobile as well as “play” are issues to consider as well,
- Underutilized talents and facilities;
- No way to ground social networking and web 2.0 tools; •
- Social isolation,
- Community. Online courses may not be able to replicate the vibrant intellectual and social community fostered by in-person education,
- Instructor workload,

- Student support - some students, including those with disabilities, may struggle to use online tools and will likely need technological support.

- Access –some students have limited or no access to computers, the Internet, and/or assistive technology.

Various types of e-learning can be distinguished. In terms of communication and learning style there are two types of e-learning synchronous and asynchronous. In synchronous instruction the teacher and students meet at the same time. In face to face instruction this means that everyone is in the same room at the same time. In online instruction synchronous instruction occurs through the use of technologies such as chat, two-way video conferencing, or audio conferencing. Online instruction is more likely to be asynchronous allowing students to access and participate in the course when they choose [14].

Synchronous e-learning is defined as Computer-assisted training where the instructor and participants are involved in the course, class or lesson at the same time (synchronized).

Asynchronous e-learning refers to learning materials that the learner can use whenever and wherever he or she wants. It connotes “on-demand” e-learning; e-learning that the learner can use when needed or when time is available [14].

There are also other types of e-learning such as mobile learning, blog learning and game learning.

Mobile learning is defined as the delivery of training by means of mobile devices such as mobile phones, PDAs and digital audio players, as well as digital cameras and voice recorders, pen scanners, etc. [3]. The potential of blogs as learning spaces for students in the higher education sector was presented by Williams and Jacobs [4]. Model g-learning was considered by Schwabe and Göth, who describe the design of the MobileGame prototype, exploring the opportunities to support learning through an orientation game in a university setting [5].

The evolution of e-learning (e-learning 1.0, e-learning 2.0 and e-learning 3.0) is related to the three generations of the Web (Web 1.0, Web 2.0 and Web 3.0).

With the advent of the Web, the major change was to have content available online. In this direct-transfer model, the instructor is the distributor of learning material in a media-rich way and addresses learners through various communication channels. This era is usually referred to as e-learning 1.0 [6].

The use of Web 2.0 technologies for teaching and learning is describing as e-learning 2.0. Web 2.0 is defined as “a space that allows anyone to create and share information online – a space for collaboration, conversation, and interaction; a space that is highly dynamic, flexible, and adaptable [15]. Web 2.0 and the associated technologies such as: wikis, blogs, podcasts, and other social web tools are well established and accepted by the students and the prevalence of these in e-learning is common. E-learning 2.0 is a collaborative model where knowledge may be socially constructed and communication is multi-directional.

The main features of the Web 3.0 technologies which differentiate it from its earlier generation, Web 2.0 are given as follows: semantic Web, openness and interoperability, global

repository of data, 3D virtualization, collaborative intelligent filtering, increased and reliable data storage capacity, higher screen resolutions, multi gesture devices and 3D touch user interface, Cloud Computing and intelligent agent systems.

One of the big things of e-learning 3.0 will be the ubiquitous access to learning resources with the use of mobile devices to virtually access anything, anytime and anywhere. [16]. Personalization is another very important trend. Personalization is seen as the key approach to handle the plethora of information in today's knowledge-based society." [17]

The usage of educational technology started from ICT education spreads into e-learning, m-learning, e-learning 2.0, e-learning 3.0 and SMART learning as the development of technology. A history of development of e-learning is presented in figure 1.

	ITC in education	e-learning	m-learning	Social learning (e-learning 2.0)	SMART learning (e-learning 3.0)
Features	Computer Assisted Instruction (CAI) Web Based Instruction (WBI)	Learning Management System (LMS)	Mobile – learning (m-learning)	Just in time learning	Intelligent adaptive learning
Main services	Cyber textbook	Cyber home study EBS Internet broadcasting	Mobile contents, Augmented Reality	App service	Online grade, smart textbook Individual portfolio
Main devices	Desktop PC	Internet PC	Mobile notebook PDA, PMP	Smart phone, Smart TV	Smart device
Time	1996	2003 -	2005 -	2010 -	2012 -

Fig. 1. A history of development
 Source: Own elaboration based on [18]

'SMART' in SMART learning means that self-directed(S), motivated (M), adaptive (A), resource free (R), technology embedded (T) education [18]. It focuses on activating online education with digital contents using smart devices. Noh, Ju, & Jung (2011) defined SMART learning as learner initiated learning which has various materials for learning and supports learner-teacher interaction [19].

Development of new solutions using intelligent agents technologies is important for the further evolution of e-learning 3.0 (smart learning).

III. CHARACTERISTIC AND PROPERTIES OF AGENT

In the literature of the subject, there are many definitions of software agents, which emphasize various features of this software. An agent can be defined as "An encapsulated computer system that is situated in some environment and that is capable of flexible, autonomous action in that environment in order to meet its design objectives." [20]. M. Woda and P. Michalec, describe that: "Agent is a process which operates in the background and performs activities when specific events occur" [21].

Software agents are attributed with a number of properties that clearly distinguish them from other types of software. Agents' characteristic is that, they act on behalf of others. Agent can be a delegate to a user, a program, another agent and it performs its tasks on behalf of them.

Agents are capable of relieving human intervention significantly and help in proper functioning of the system. From the various characteristics of agents the most frequently mentioned are: [22] [20] [23]:

1. **Autonomy:** Autonomy corresponds to the independence of a party to act as it pleases. Autonomous agents have control both over their internal state and over their own behavior.
2. **Reactive (sensing and acting):** the agent responds based on the input it received and according to the environment. It responds in timely fashion to changes in the environment.
3. **Proactive:** A proactive agent is one that can act without any external prompts. It acts in anticipation of the future goals
4. **Flexibility:** the agents are dynamic as their reaction is dynamic and varies according to the environment. Actions are not scripted.
5. **Communication:** It can be defined as those interactions that preserve the autonomy of the parties concerned. Communicates with its user and other agents.
6. **Mobility:** it is important for the agents to be able to move to other location (machine or environment) and to continue their tasks there.
7. **Temporally continuous.** It is continuously running process.
8. **Learning.** Changes its behavior based on its previous experience.

Agent software can be classified according to their functionality. Examples of very diverse agent's activities are presented below:

Agent supporting user. Its task is to help the user to use applications, devices or websites. This type of software often gives the impression of a contact with a real person.

Agent as an assistant which role is for example to manage the calendar of meetings or to search some information online according to the user's interests.

Email agent that performs the initial selection of e-mail which include spam rejection, sorting e-mails, checking for viruses, prioritizing messages.

Agent that search the Internet resources in order to gather information that is potentially necessary to the user.

Agent that manages network, supervises computer networks, detects failures and responses to threats. It also monitors networks and creates statistics.

From this wide selection of agent software, for the purpose of these considerations, the most important solutions will be those that can be used in e-learning.

Dynamic development of specialized software agents has stimulated the creation of multi-agent systems. The direction of changes exposes the skills of communication and cooperation of specialized agents. Important matters of multi-agents software include division of tasks between agents, selection of communication method, interaction between agents, protocols and system architecture. E-learning is an area in which the effective solutions of specialized autonomous agents and multi-functional multi-agent systems are expected.

IV. THE CONCEPT OF A MULTI-AGENT INTELLIGENT SYSTEM FOT E-LEARNING

Multi-agent systems are computer system that use the agent software. Agents can offer various services, customized to the needs of both students and teachers.

Agents as assistants can support the distance learning process. Student's assistant can be used to search educational materials and to monitor the user's progress in the online course. Teacher's assistant can help to distribute the course materials among students and can observe the learning progress of students.

Various roles of agents in e-learning systems were identified, therefore, only the concept of a multi-agent system can be used to fulfill such a complex task. The concept of an intelligent multi-agent system for the distance learning is presented in Figure 2.

The environment of the system includes agents' management system. The second essential element is the channel of communication between agents which is used to exchange information between agents. The communication between agents is carried out by sending messages using the standard ACL (Agent Communication Language). Figure 2 highlights a catalog of services provided by software agents.

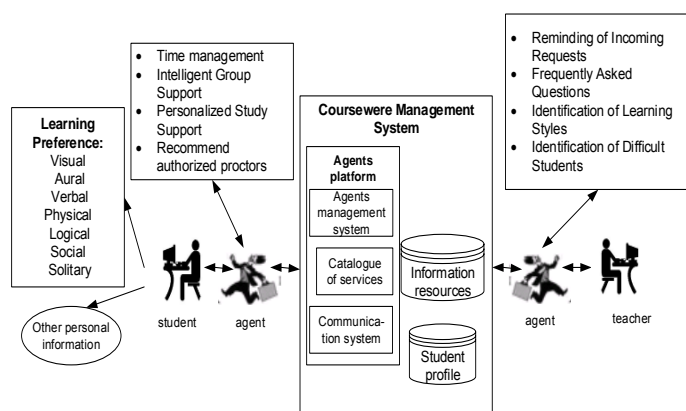


Fig. 2 The concept of an intelligent multi-agent system for distance learning
 Source: Own elaboration based on [24]

Everybody learns in a different way, therefore when possible people try to adapt the most suitable learning style to their needs. One of the services offered by the multi-agents software is identification of students' learning style. In literature there are many typologies of styles of learning. One of the most popular is Memletics Learning Styles that differs seven basic methods [25].

- Visual (spatial). Student prefers using pictures, images, and spatial understanding.
- Aural (auditory-musical). Student prefers using sound and music.
- Verbal (linguistic). Student prefers using words, both in speech and writing.
- Physical (kinesthetic). Student prefers using your body, hands and sense of touch.
- Logical (mathematical). Student prefers using logic, reasoning and systems.
- Social (interpersonal). Student prefers to learn in groups or with other people.
- Solitary (intrapersonal). Student prefers to work alone and use self-study.

TABLE I. EXAMPLES OF SOFTWARE AGENTS USAGE AS ASSISTANTS TO TEACHER AND STUDENT IN THE DISTANCE LEARNING

Functionality	Description
Student support	
Time management	The agent helps students to manage their time effectively by notifying students of due dates of assignments and appointments and develops right progress schedules based on students' activity schedules
Intelligent group support	The agent helps students by organizing study groups based on their study interests to achieve the best learning performances
Personalized study support	The intelligent agent recommends right learning styles to students based on their learning preferences to improve their learning effectiveness
Recommend authorized proctors for students to take exams	The intelligent agent searches the nearby proctor centers to find the best ones where each student can take his or her exam
Teacher support	
Reminding instructors of incoming requests	Through short messages on the cell phone, intelligent distance education systems alert instructors that new questions are awaiting responses.
Frequent asked questions (FAQ) management	The intelligent system builds a frequently asked questions case base for instructors to retrieve and adapt in the future
Identification of student's learning styles	The intelligent agent helps to identify student learning styles, helps instructors develop right teaching strategies, and offers personalized suggestions.
Identification of difficult	The intelligent systems help instructors to identify students with difficulties learning

students	in order to develop personalized teaching strategies.
----------	---

Source: Own elaboration based on [23]

students) properties such as reactivity, proactivity and learning are very important.

V DEVELOPMENT TRENDS OF SOFTWARE AGENTS

Most important properties of software agents have been enumerated in Tables 2 and 3, while in the columns services offered by agents for students and teachers were mentioned. The meaning of each property was estimated in comparison to the effective realization of the service.

In order to evaluate the properties of software agents a team of 5 experts was appointed (4 co-workers of the author and the author). Each expert has a degree in computer science, a certificate of training in the field of e-learning teaching and has organized e-learning classes. Every expert made an independent assessment. Basing on the collected assessments the final mark was given.

Three levels scale was taken into account: very important property (VIP), less important property (LIP) and meaningless property (MP).

TABLE II. PROPERTIES AND SERVICES TO BE DEVELOPED IN THE FUTURE OF SOFTWARE AGENTS SUPPORTING STUDENTS

Properties of software agents	Services offered by agents for students			
	Time management	Intelligent group support	Personalized study support	Recommend authorized proctors for students to take exams
Autonomy	MP	VIP	LIP	LIP
Reactivity	LIP	VIP	VIP	VIP
Proactivity	VIP	VIP	VIP	VIP
Flexibility	MP	LIP	LIP	LIP
Communication	VIP	VIP	VIP	LIP
Mobility	MP	MP	MP	MP
Temporally continuous	VIP	LIP	VIP	LIP
Learning	MP	VIP	VIP	VIP

Source: Own elaboration

It turns out that the services offered by software agents such as time management or reminding instructors of incoming requests are implemented efficiently enough therefore there is no need to further developed agents for this services (the largest number of ratings as meaningless property MP). For the improvement of identification services (recommend authorized proctors for students to take exams, identification of student's learning styles and identification of difficult

TABLE III. PROPERTIES AND SERVICES TO BE DEVELOPED IN THE FUTURE OF SOFTWARE AGENTS SUPPORTING TEACHERS

Properties of software agents	Services offered by agents for teachers			
	Reminding instructors of incoming requests	Frequent asked questions (FAQ) management	Identification of student's learning styles	Identification of difficult students
Autonomy	MP	MP	LIP	VIP
Reactivity	LIP	LIP	VIP	VIP
Proactivity	VIP	VIP	VIP	VIP
Flexibility	MF	LIP	LIP	VIP
Communication	VIP	VIP	LIP	LIP
Mobility	MP	LIP	MP	MP
Temporally continuous	VIP	VIP	LIP	VIP
Learning	MP	VIP	VIP	VIP

Source: Own elaboration

CONCLUSIONS

Identification of the desirable properties and services of every specialized group of software agents supporting students and teachers will help to target the research and improve the existing solutions in this area.

The research shown that for both, agents supporting students and agents assisting teachers, the most desired properties, which should be develop in the future, are proactivity and communications.

The development of the learning property of the agent is valued as very important for six out of eight offered services. It is only valued as meaningless for time management and reminding instructors of incoming requests.

Mobility understood as being able to transfer itself from one machine to another is not a property of agent that is important for e-learning. Out of eight services, only in the case of FAQ Management it was evaluated as with little importance, while the remaining seven were evaluated as meaningless.

It was estimated that the flexibility in the development of software agents in e-learning is rather of a small importance. Only in the case of identification of difficult students it was rated as very important, because depending on the currently presented subject, a group of students who have difficulties is being created

E-learning is a teaching method that is gaining a growing number of supporters among both teachers and students. For

the further development of e-learning, eg. in the direction of learning 3.0 it is essential to effectively adopt new solutions, such as software agent technology supporting both students and teachers.

REFERENCES

- [1] D. Jelonek, A. Nowicki, and L. Ziora, "The Application of e-Learning in the Didactic Process at the Faculty of Management in Czestochowa University of Technology. Organization. Tools. Model", Proceedings of Informing Science & IT Education Conference (InSITE), pp.143-156, 2014.
<http://proceedings.informingscience.org/InSITE2014/InSITE14p143-156Jelonek0467.pdf> (2014-12-20).
- [2] C. D Dziuban, J.L. Hartman, and P.D. Moskal, "Blended learning", ECAR Research Bulletin, 7. Retrieved April 27, 2008 from <http://net.educause.edu/ir/library/pdf/erb0407.pdf> (2014-12-10)
- [3] D. Keegan, "The future of learning: From elearning to mlearning", Fern Universitat –Hagen, November 2002.
- [4] J. B. Williams, and J. Jacobs, "Exploring the use of blogs as learning spaces in the higher education sector". Australasian Journal of Educational Technology, 2004, vol. 20(2), pp. 232-247.
- [5] G. Schwabe, and Ch. Göth, "Mobile learning with a mobile game: Design and motivational effects", Journal of Computer Assisted Learning, 2005, vol. 21(3), pp. 204–216.
- [6] F. Hussain, "E-LEARNING 3.0 = E-LEARNING 2.0 + WEB 3.0?", IADIS International Conference on Cognition and Exploratory Learning in Digital Age (CELDA 2012), <http://files.eric.ed.gov/fulltext/ED542649.pdf> (2015-01-05).
- [7] S.-L. Huang, and J.-H. Shiu, "A User-Centric Adaptive Learning System for E-Learning 2.0." Educational Technology & Society, 2012, 15 (3), pp. 214–225.
- [8] C. Safran, D. Helic, and C. Gütl, "E-learning practices and web 2.0", Paper presented at the 2007 International Conference on Interactive Computer Aided Learning, Villach, Austria, September 2007.
- [9] F. Alonso, G. López, D. Manrique, and J.M. Viñes, "An instructional model for web-based e-learning education with a blended learning process approach", British Journal of Educational Technology, 2005, 36(2), pp. 217-235.
- [10] S. Guri-Rosenblit, "Distance education and e-learning: Not the same thing", Higher Education, 2005, Vol 49(4), pp. 467-493.
- [11] B. Nikolić, and L. Ružić-Dimitrijević, "Distance Learning – from Idea to Realization", Proceedings of Informing Science & IT Education Conference (InSITE) 2010.
<http://proceedings.informingscience.org/InSITE2010/InSITE10p369-384Nikolic806.pdf> (2015-01-02).
- [12] V. Nedeva, E. Dimova, and S. Dineva, "Overcome Disadvantages of E-Learning for Training English as Foreign Language", The 5th International Conference on Virtual Learning ICVL 2010 (2015-01-10).
- [13] Challenges and Disadvantages of E-learning and Distance Learning, 2009, <http://compassioninpolitics.wordpress.com/2009/09/26/>
- [14] N. Al-Taie, "The Effect of Using E-Learning Curriculum And Traditional Classroom Curriculum: Comparison &Merits", ICIT 2013 The 6th International Conference on Information Technology, <http://icit.zuj.edu.jo/icit13/index.html> (2015-01-10).
- [15] K.A. Coombs, "Building a library web site on the pillars of Web 2.0", Computers in Libraries, 2007, Vol. 27 No. 1, available at: www.infoday.com/cilmag/jan07/Combs.shtml (2014-12-18).
- [16] D. Baird, "Learning 3.0: Mobile, Mobile, Mobile Barking Robot", 2007, Retrieved March 21, 2012
http://www.debaird.net/blendededunet/2007/02/learning_30_mob.html (2014-12-16)
- [17] M. Ebner, S. Schön, B. Taraghi, H. Drachler, and P. Tsang, "First steps towards an integration of a Personal Learning Environment at university level", In R. Kwan et al. (Eds.), ICT 2011, CCIS 177 (pp. 22–36), Springer-Verlag Berlin.
- [18] K. Soon-Hwa, S. Ki-SSang, and P. Se-Young, "Exploring the Technological Factors in SMART Learning Affecting Creativity, Advances in Educational Technologies", Proceedings of the 2014 International Conference on Education and Modern Educational Technologies (EMET 2014, Edited by N. Mastorakis, P. Dondon, P. Borne, Santorini Island, Greece, July 18-20, 2014, pp.161-166.
- [19] Noh, Ju, & Jung, "The conditions of SMART Learning", digital policy study, 2011, 9(2), 79-88.
- [20] N. Sivakumar, K. Vivekanandan, B. Arthi, S.Sandhya, and V. Katta, "Incorporating Agent Technology for Enhancing the Effectiveness of E-learning System", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, May 2011, ISSN (Online): 1694-0814, www.IJCSI.org 4.
- [21] M. Woda, P. Michalec, "Distance Learning System: Multi-Agent Approach", Journal of Digital Information Management-September 2005.
- [22] D. Jelonek, and A. Chluski, "The role of software agents in distance education (Rola agentów programowych w nauczaniu na odległość)", Business Informatics, 2011, No 17, pp. 86-93.
- [23] E. Kaasinen, "Usability Challenges in Agent Based Services", in: H. Zuidweg(ed.), Intelligence in Services and Networks. Paving the Way for an Open Service Market: 6th International Conference on Intelligence and Services in Networks, IS&N'99, Barcelona, Spain, April 27-29, 1999, Proceedings.
- [24] L. Xiaoqing, "Intelligent agent – supported online education", Decision Science Journal of Innovative Education, 2007, Vol.5, No.2, pp.311-331
- [25] Memletics Learning Styles Inventory. Free publication provided by www.memletics.com; <http://www.crs.sk/storage/memletics-learning-styles-inventory.pdf> (2015-01-06).

Privacy Policy of E-Government Websites and the Effect on Users' Privacy

Maryam Al-Jamal & Emad Abu-Shanab

MIS Department, IT College, Yarmouk University, Irbid, Jordan,
maryam.aljamal@yahoo.com & abushanab@yu.edu.jo

Abstract—The rapid developments in information and communication technologies require citizens to provide more information to their government within the context of e-government. Providing personal information to e-government websites is necessary to benefit from its services, where providing such information over the Internet raises privacy concerns by users. This paper argues that a privacy policy can be a good guarantee for users' privacy and a factor that supports the intention to use e-government services. The literature and international reports were explored to understand the issues related to privacy policy in e-government and their importance to users. Principle by the OECD and FTC are widely used by researchers to set the stage for better development of privacy policies. Finally, two frameworks were proposed to guide future research and recognize the factors affecting the adoption process of e-government.

Keywords—E-government, Privacy, Privacy policy, Security, FTC principles, OECD principles, proposed framework.

I. INTRODUCTION

The nature of interaction with information technology requires users to provide more information about them. Internet has the most significant effect on our lives; since we use it for communication, learning, entertainment, business applications and many others. To benefit from all previously mentioned applications, users are required to disclose their private information. Providing such information over the Internet raises many concerns for the users. Privacy of users' information is one of these concerns.

One of the major applications of information and communication technology (ICT) and the Internet is e-government. E-government websites provide us with information and services. The amount of data and information gathered by governments' websites is increasing, and users don't know the extent to which his/her information is secure and protected. The presence of a privacy policy is required in e-government websites to ensure users' privacy.

Although privacy policy can be a guarantee for citizens' data protection on e-government websites, there are still some websites that don't adhere to such provision. Even privacy policy can't be that much effective if there are no privacy protection laws in the country, or no clear definition of privacy policy or what it should contain.

Research indicated that trust in e-government websites is a significant predictor of their adoption of such websites and the intentions to use those [1]. Users' trust in e-government websites can be affected by the presence of privacy policy on their websites. Still there are well crafted privacy policies and deficient ones.

This study will explore the literature to understand the privacy issues related to e-government websites and

information systems. Privacy issues revolve around privacy policy, its definition, its importance, and its contents. Also, the factors affecting users' attitudes toward privacy policies will be explored, where we present some globally known principles for developing privacy policies. The presence of privacy policy in e-governments websites will be investigated along with the quality of these policies. The concept of privacy and its implications on e-government is important. Privacy protection solutions will also be presented.

The structure of this paper is as follows: a literature review of previous related work will be presented in an overview of e-government, followed by an exploration of privacy issues in e-government websites. Finally, an investigation of privacy policy in e-government will be conducted to conclude to a conceptual framework to guide our future research. Conclusions and future research will be depicted at the end.

II. LITERATURE REVIEW

Using e-government websites and interacting with its systems and interfaces is the major indicator of such projects success. To enhance the chances of citizens' use of e-government websites, privacy of information should be maintained. The following section will explore the literature related to such issue.

A. Overview of e-government

Many researchers have defined e-government in many several ways. E-government revolves around using the Internet to provide services to citizens, businesses and employees to enhance efficiency and effectiveness of private and public sectors [2]. Others have considered e-government as "the use of information and communication technology (ICT) and particularly the Internet to deliver information and services by the

government to its customers (businesses and citizens)” [1, p. 39].

In addition to that, e-government is defined as “the wide and efficient use of application of different technologies by governmental departments and ministries to connect with and better serve the citizens” [3, p. 278]. For the purpose of this study, and based on the previous definitions explored, we can define e-government as: utilizing ICT tools and applications to provide service and information to citizens, businesses and public employees in a better, more efficient and effective way that protect their privacy of information.

As widely recognized, e-government has many benefits to citizens, businesses and government itself. It’s noteworthy to citizens that improving accessibility to public services is important, but enhancing transparency and the effectiveness of government performance is more important. For businesses e-government is suitable, fast and cost-effective for getting the needed information and services. For governments, e-government is an innovative tool that allows governments to know the needs of their people and serve them quickly and at a reduced cost [4].

Generally e-government helps in fighting corruption and bribes [5], reforming the social and economical status in the country, enhancing governance, saving time in providing services and offering the citizens a higher accessibility to policies, standards, laws and information [3]. Also, e-government has been known as an effective tool for increasing accountability, enhancing transparency rates and fostering e-democracy [6] [7].

Besides all previously mentioned benefits of e-government, there are some obstacles and challenges hindering the progress of e-government projects. These challenges may be due to government agencies and their users. Many studies included the following obstacles: privacy issues, digital divide, availability, trust, security application, improper integration between systems and governmental departments, infrastructure costs, lack of legal frameworks supporting e-government and cultural issues [4] [8].

Researchers now view e-government as an effective strategic tool for administrative reform in public sector, at all levels of government bodies. Also, they considered the emergence of web 2.0 tools and development (such as: mobile devices, wikis, blogs and social media) is the reason behind enlarging the spectrum of people participating and interacting with government bodies, and even government agencies among each other. Moreover, government websites can be a great component for facilitating public information sharing [9]. So e-government has many channels to reach its stakeholders besides its websites.

B. Privacy in e-government

Moving from the traditional government to e-government resulted in a loss of privacy and security of users’ personal data; this loss was caused by shifting from centralized/closed systems to decentralized/open governance systems. Personal data is defined as any type

of data that can reveal person’s identity (directly or indirectly). Examples of personal data are: ID number or social security number SSN, employment number, age and religion [10].

Privacy is a broad term that is defined in many ways depending on the context, environment or perspective. However, privacy is the state when an individual can control personal information about his/her self and how, why, what and who knows such information. Other concepts regarding privacy are related to e-government nature; that is “online privacy”. Simply, online privacy is person’s privacy over the Internet [11].

It is important here to clarify the relationship between privacy and security. It is noticed in the literature that privacy and security are always mentioned together and even explored together within the e-government literature [12]. The reason behind that is that security is known as protecting the system against threats like hackers, crackers and viruses. These risks threaten the privacy of systems’ users. Then security of the system is the gate to invading privacy of information. So e-government websites need to grant the needed level of privacy along with the security mechanisms intended to be used [13] [14].

The literature of e-business and e-commerce frequently mentioned a situation when the website sells/gives information of users to a third-party, as a threat to privacy. Such situation influences citizens’ trust in e-government. Trust in e-government is explored in the literature based on two dimensions: trust in the technology and trust in the government itself [1]. Trust in technology can be solved by enhancing the security levels and the legal framework related to online service. Trust in government is the responsibility of the government itself to improve its image. In the second situation trust is gained in an ongoing process [1].

E-government can bring to societies and public systems more efficiency in offering services, higher accessibility to public services, empowered participation, and better transparency. Public participation is widely recognized for playing an important role in improving government activities and communication with citizens [15]. So information that is provided by citizens through e-government websites should be secured and their privacy must be preserved by the government [4].

As e-government is shifting to open government, more emphasis is put on transparency and information exchange. The more countries are embracing e-government, the more they are enhancing the transparency of their systems [7]. Theoretically speaking, the more transparent organizations are becoming, the more they slip into the trap of violating privacy issues. In some situations organizational transparency is reduced to protect others’ rights like privacy [6]. Transparency must be balanced against privacy in a way that adheres with societal norms and without violating international standards.

Research also focused on the adoption process of e-government services where some researchers concluded to a set of factors affecting user’s intention to use e-

government websites like: system quality, service quality, information quality [16], risk perceptions [12], trust, perceived ease of use, perceived usefulness, and social influence [17]. Others examined several factors that affected the use of SMS based e-government services in Jordan; they found that “perceived risk to users’ privacy” has been ranked the fourth among many other factors in predicting the adoption process which indicates the importance of privacy and security [18].

In a study profiling non-users of e-government, it is found that the negative attitude to e-government services was not the reason behind not using the services. Researchers argued that it may be a result of perceived risks (security and privacy risks) in using e-government services. Governments need to pay attention to the importance of high security techniques used in their systems to protect their systems and people’s privacy [19]. Such efforts will improve government’s reputation and citizens’ intention to use its systems.

Users’ satisfaction is measured by how frequent they use the service and visit the website again and again. Irani et al. [20] concluded that citizens’ satisfaction and trust in e-government are increased when it provides them with secured and privacy oriented systems. So citizens’ online privacy must be guaranteed as it is a crucial factor for e-government’s success.

A study of citizens’ e-government preferences concluded to four segments of users: risk-conscious, balanced, recourse-conservative and usability-focused. The level of privacy that citizens require was affected by the type of service they use. In the same study it is found that citizens’ concerns were greater when they filed their taxes online compared to online appointment booking [21].

As mentioned previously, in their pursue to open communication with their citizens and reach them where ever they are, governments utilized social media and tried to benefit from such channel [22]. Using social media channels, many obstacles and threats are facing governments and e-government projects. Examples of these threats are: lack of government possibility to ensure users’ privacy, lower control on social media contents, and the absence of legal framework governing activities in social media [23]. It is obvious that the threat level on users’ privacy in e-government is related to the channel used by the government.

Regarding privacy protection solutions, there are technical solutions and legal-based solutions. Some researchers asserted that the need for laws to ensure privacy of information is more important. They emphasized the importance of issuing the needed laws and enforcing them. More over e-government has been considered as a tool for pushing forward e-business movement by setting such laws [13].

In an analysis of European Union countries’ policies regarding privacy and security; it is found that there is a difference in how each country understands privacy and security issues [24]. The author proclaimed that they shared the assumptions about policy formulation and the need of governmental intervention in the policy

formulation process. Such conclusion indicates the difficulty of formulating and issuing the needed laws, standards and policies related to e-government environment.

Many Studies concluded that parties administering e-government projects (mainly the government itself) should develop information security goals and make sure that resources are available to achieve these goals. Surly, an investment in security techniques and mechanisms must be established and developed to improve the security and privacy status. That’s because users’ privacy concerns play a significant role in affecting e-government performance [14].

C. Privacy policy in e-government

Although users are concerned about their privacy over the Internet they have fair knowledge on how to protect their privacy. Users try to protect their privacy by deleting cookies, being more conservative in providing unnecessary information. There are specialized providers of privacy seals and standards like TRUSTe, PriceWaterhouseCoopers PWC, BBB Online and WebTrust [13]. This study will focus on privacy policy as a major privacy assurance tool.

Privacy policy can be defined as “*legal document that defines how the website gathers information from the user and how it uses this information*” [25, p. 88]. Privacy policy clarifies for website users how their data is being collected, the purpose of data collection, and the different uses of such data. Researchers concluded that culture is a significant factor affecting users’ attitudes toward the content of a privacy policy. A group of researchers compared the responses of Russian and Taiwanese users in regard to information provided online; they found that Taiwanese trust has increased when they knew that their information is secured, while Russians trust didn’t increase [26].

A study conducted in China found that most websites have a type of privacy discloser. In the same study, it is noticed that the majority of websites collect ID number/SSN; they interpret it as a step by the government to protect their citizens’ from fraud. Such step might be considered in other cultures as an intrusion of privacy [27]. Based on the previous two studies, we can infer that the needed privacy level is affected by the difference of cultural perspectives.

Contents of privacy policy mainly depend on laws enforced in the country regarding this issue and the requirements of the organizations interacting with users. Privacy policy should clarify what data is collected from users, why it is collected and how it will be used. Moreover, privacy policy must be readable and understandable by all of the targeted users of the website [28]. The authors conducted a study regarding the Saudi e-government websites, 28% of websites had privacy policy, while the other 72% did show any kind of privacy policy or agreement. On the other hand, among the websites that have privacy policy 60% of them have a well formulated privacy policy and 40% have weak ones. We can infer from the previous study that the

presence of privacy policy is not enough, the quality of privacy policy must also be considered.

A proposed framework by Jha and Bose [10] tried to set some set of standards for planning privacy policy; the framework “CCAGM” stands for: centralization, characterization, access gating and monitoring. It is claimed that the framework can be an effective tool for administering security and privacy issues within the e-government context. *Centralization* means storing all data and records in one secure location, and that is intended to prevent duplication and scattering data in locations that might be unsecured. *Characterization* means that data will be classified as private, public or personal. And *Access Gating* is the mechanism of controlling access to data from different users, like password or SSN. *Monitoring* is to monitor various transactions and check standards formed by the central authorities.

Another important issue regarding privacy policies formulation is the frequent changes of privacy policy. A study presented the problem of privacy policy within Facebook context and considered it as a misleading one due to its frequent changing nature. The frequent changes of privacy policies confused the users of the website about what information they are sharing, to whom and how their information is used. Regarding this issue the federal trade commission (FTC) threatened Facebook to take an action against them (as a regulation body), then Facebook reached a settlement to make its privacy policy consistent and transparent [29]. The question of whether government privacy policies are changing frequently emphasizes the importance of governing laws regarding formulating privacy policy.

D. Principles for developing privacy policy

The context, conditions and guidelines for building a privacy policy are researched by non-academic parties, where some institutions consider themselves guardians for the privacy of citizens’ data. The Federal Trade Commission (FTC) is one example, and the Organization for Economic Cooperation and Development (OECD) is another example of organizations that have set principles and standards for writing and developing privacy policies. Some researchers considered the FTC principles as more flexible and realistic framework to guide such process [26].

The OECD included eight main privacy principles. These principles are: Safe guard, collection limitation, data quality, purpose specification, use limitation, openness, individual participation and accountability [30]. Safe guard means that data should be protected and secured from any unauthorized access or risks. Collection limitation means that there should be limits for data collection and data should be collected in a legal manner. Data quality means that data should be accurate, up-to-date, complete and relevant to the purpose of collection. Purpose specification reflects the purpose behind collecting the data and must be stated to users (to take their consent) before the collection process starts and whenever the purpose has changed. Use limitation:

personal data should not be revealed or used for other purposes than originally intended, unless the user is informed or the law permits. Openness principle means that organizations must make privacy policy explaining their policies regarding data collection and management [31].

Individual participation means that users must have the right to get their data from data collectors, citizens should have open communication with data collectors at any stage, know the reasons behind any denied requests, and to have control over their data (deletion and change). Accountability means that the service provider is responsible for enforcing and adhering to all other OECD principles applied in their system/website [30].

The FTC contains five main privacy principles. These principles are: Notice, choice, access, security and enforcement. Many researchers used these principles in evaluating privacy policies [26] [28]. Notice means that the system/website must explain clearly what data it collects, why and how will be used. Choice: the website should inform users if they will give their data to a third party and why, and must clearly ask for the users’ permission. Access: the website should allow users to review, correct or delete personal information collected by the website. Security principle means that any unauthorized access to users’ data must be prevented and the highest security mechanisms must be used and applied to protect users’ personal data. Enforcement: the website states that there is a law governing any violations of privacy and the website will take actions against the violators according to the stated law [26]. The following table summarizes both OECD and FTC principles, where we matched both set of principles against each other in a proposition for researchers and to guard against redundancy of issues.

Table 1: Matching principles from OECD & FTC

OECD principles	FTC principles	Matched principles
1. Safe guard	A. Notice	(A,4,2)
2. Collection limitation	B. Choice	(B,5)
3. Data quality	C. Access	(C,7, 5)
4. Purpose specification	D. Security	(D,1)
5. Use limitation	E. Enforcement	(E,8,6)
6. Openness		
7. Individual participation		
8. Accountability		

From Table 1 it is noticed that “data quality” principle of OECD didn’t match with any principle of FTC principles. That may indicate that OECD principles are more comprehensive than FTC principles. However, Wu, Huang, Yen and Popova [26] used the FTC principles for judging privacy policies because these principles are more flexible and realistic, and they are more oriented to users and risks associated with personal data collection [26, p. 891]. Both OECD and FTC

principles are valid principles and widely recognized ones.

III. PROPOSED RESEARCH FRAMEWORK

As e-government phenomenon is spreading worldwide due to the great contributions of ICT tools, more personal information is exchanged between governments and their citizens. This necessitates posting and enforcing privacy policies on e-government websites. Many studies were conducted on privacy policies in e-commerce websites, and some on e-government websites. To understand the evolving domain of e-government and to understand the factors influencing its success, we have proposed a framework that aligns the directions of citizens with those of the government. The main question is “Does the existence of a privacy policy on e-government website influences users’ concerns about their privacy?”

Based on the previous discussion in literature review section, Figure 1 shows a proposed framework for citizens-government relationship regarding privacy. In this framework users and governments are the main players in e-government. When users have the intention to use e-government websites, there are three main important factors that influence their attitudes: security, privacy and trust. On the other hand, governments need to encourage people to use e-government websites (as a measure of success), where they must provide three basic things: enforced laws, privacy policy and security mechanisms. Security leads to more controlled privacy which both facilitates more trust. Corners of the figure represent the most important components: users, government and *privacy dimensions*. Also, in the figure

security, laws and trust are supporting components leading to the heart of the relationship “e-government”.

The Citizen-Government Privacy Alignment Model (CGPAM) reflects on major issues previously explored in the literature: privacy, security and trust. The technology adoption literature is rich with studies that investigated the perceptual or attitudinal predictors of e-government success. Factors like perceived usefulness, perceived ease of use, social influence, perceived facilitating condition, and many other factors, were all explored in the context of e-government websites. This framework looks from a different angle to the phenomenon, where privacy policy, its contents, and its enforcement perceptions are major predictors of e-government success.

To connect the major stream of research into this framework another framework is proposed which tries to differentiate between the major predictors of use in the literature (perceived usefulness, perceived ease of use and social influence) and what we are concerned about here (security, privacy and trust). We are hypothesizing here that citizens will use e-government websites and services in a shallow mode (simple information, not critical, not financial, and not important) when they feel that security, privacy and trust in jeopardized. While citizens will start an in-depth, real use of e-government services when they feel it is secured, has a privacy policy, and they trust it. Such use will lead to the continuous use, which is the ultimate objective of e-government projects. Figure 2 depicts the Continuous E-government Use Model (CEGUM).

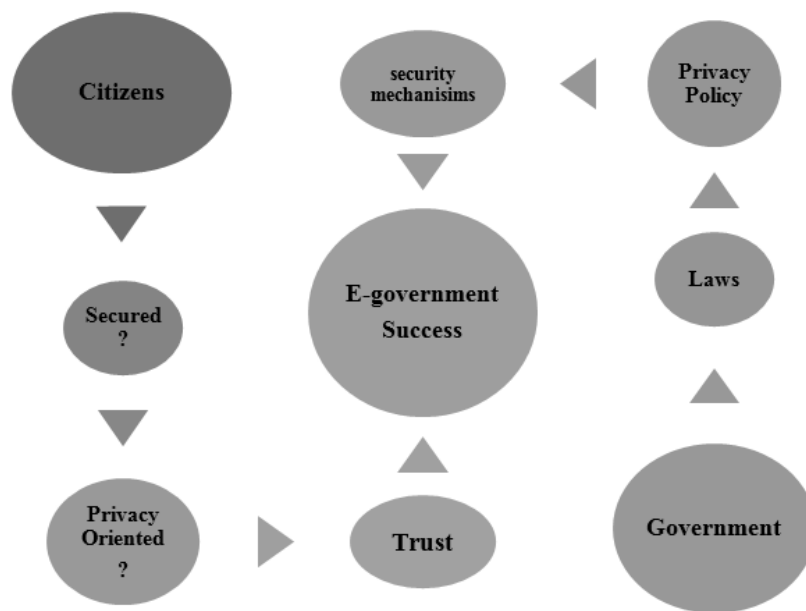


Figure 1: The Citizen-Government Privacy Alignment Model (CGPAM)

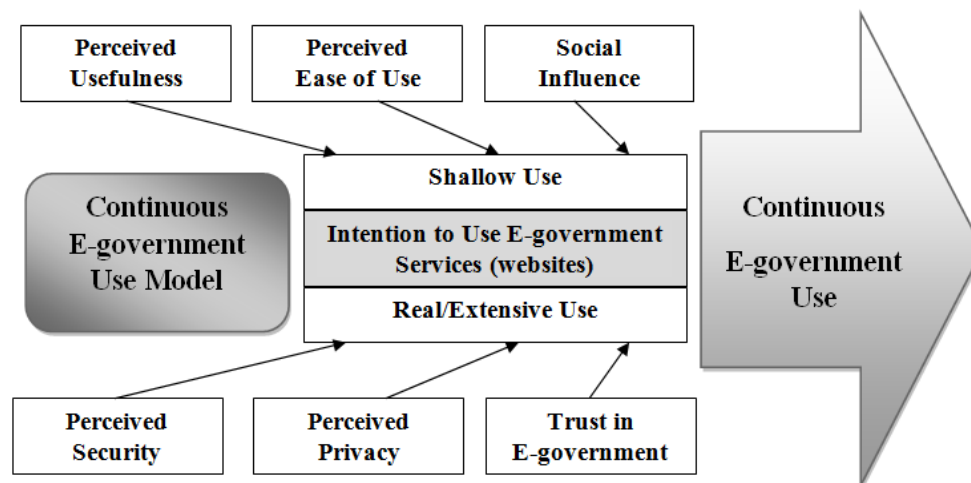


Figure 2: The Continuous E-government Use Model (CEGUM)

IV. CONCLUSION

This paper reviewed the literature to understand the issues related to privacy in e-government and its context. Research and reports asserted the importance of privacy in e-government and its effect on trust and adoption of e-government initiatives among citizens and businesses. Privacy has a significant effect on government performance and users' satisfaction. Privacy level needed is affected by the service being used by the user, and its preservation is achieved by applying high security techniques and enforcing solid laws and regulations. The existence of a privacy policy to aid in defining the relationship between government and users is also significantly important. Privacy policy is a legal document, where users' attitude toward it is affected by their culture. There are many widely known principles for writing privacy policies; the famous ones are visited in this study and they are the FTC and OECD principles. However, privacy policy existence is not an enough indicator for protecting users' privacy; its quality must also be considered. Also, trust in e-government is a key factor that determines users' beliefs and intentions to use e-government services.

This study proposed two frameworks to guide future research; the first is the Citizen-Government Privacy Alignment Model (CGPAM), which focuses on the alignment between citizens' concerns and government policies and processes. The second framework sums the factors that contribute to continuous use of e-government services/websites. Future research is recommended to test the two frameworks and understand better the context of adopting e-government within a privacy policy context. Also, it is recommended to empirically test the second model to see if such factors are predicting the intention to use e-government and whether it will lead to continuous use.

REFERENCES

- [1] Abu-Shanab, E. & Al-Azzam, A. (2012). Trust Dimensions and the Adoption of E-Government in Jordan. *International Journal of Information Communication Technologies and Human Development*, Vol. 4(1), pp. 39-51.
- [2] Nawafleh, S. A., Obiedat, R. F. & Harfoushi, O. K. (2012). E-Government between Developed and Developing Countries. *International Journal Of Advanced Corporate Learning*, Vol. 5(1), pp. 8-13.
- [3] Kayrouz, A. & Atala, I. (2014). E-GOVERNMENT IN LEBANON. *European Scientific Journal*, Vol. 10(7), pp. 277-283.
- [4] Gajendra, S. Xi, B. & Wang, Q. (2012). E-Government: Public Participation and Ethical Issues. *Journal Of E-Governance*, Vol. 35(4), pp. 195-204.
- [5] Abu-Shanab, E., Harb, Y. & Al-Zo'bie, S. (2013). Government as an Anti-Corruption Tool: Citizens Perceptions. *International Journal of Electronic Governance*, Vol. 6(3), 2013, pp. 232-248.
- [6] Halachmi, A. & Greiling, D. (2013). Transparency, E-Government, and Accountability. *Public Performance & Management Review*, Vol. 36(4), pp. 572-584.
- [7] Abu-Shanab, E. (2013). The Relationship between Transparency and E-government: An Empirical Support. IFIP e-government conference 2013 (EGOV 2013), September 16-19, 2013, Koblenz, Germany, pp. 84-91.
- [8] Basamh, S. S., Qudaih, H. A. & Suhaimi, M. A. (2014). E-Government Implementation in the Kingdom of Saudi Arabia: An Exploratory Study on Current Practices, Obstacles & Challenges. *International Journal of Humanities and Social Science*, Vol. 4(2), pp. 296-300.
- [9] Sandoval-Almazan, R. & Gil-Garcia, J. R. (2012). Are government internet portals evolving towards more interaction, participation, and collaboration? Revisiting the rhetoric of e-government among municipalities. *Government Information Quarterly*, Vol. 29, pp. S72-S81.
- [10] Jha, A. & Bose, I. (2013). A Framework for Addressing Data Privacy Issues In E-Governance Projects. *Journal Of Information Privacy & Security*, Vol. 9(3), pp. 18-33.
- [11] Brandimarte, L., Acquisti, A. & Loewenstein, G. (2013). Misplaced confidences privacy and the control

- paradox. *Social Psychological and Personality Science*, Vol. 4(3), pp. 340-347.
- [12] Khasawneh, R., Rabayah, W. & Abu-Shanab, E. (2013). E-Government Acceptance Factors: Trust And Risk. The 6th International Conference on Information Technology (ICIT 2013), 8-10 May, 2013, Amman, Jordan, pp.1-8.
- [13] Cepani, L. (2012). The Security and Privacy Issues as One of the Barriers Impeding the E-Business Development in Albania. *Annals of the Alexandru Ioan Cuza University-Economics*, Vol. 59(1), pp. 353-362.
- [14] Zu'bi, M. H. & Al-Onizat, H. H. (2012). E-Government and Security Requirements for Information Systems and Privacy (Performance Linkage). *Journal of Management Research*, Vol. 4(4), pp. 367-375.
- [15] Al-Dalou', R. & Abu-Shanab, E. (2013). E-Participation Levels and Technologies. The 6th International Conference on Information Technology (ICIT 2013), 8-10 May, 2013, Amman, Jordan, pp.1-8.
- [16] Qutaishat, F. T. (2013). Users' Perceptions towards Website Quality and Its Effect on Intention to Use E-government Services in Jordan. *International Business Research*, Vol. 6(1), pp. 97-105.
- [17] Abu-Shanab, E. (2014). Antecedents of Trust in E-government Services: An empirical Test in Jordan. *Transforming Government: People, Process and Policy*, in press and expected to appear in 2014.
- [18] Al-ma'aitah, M., Altarawneh, M. & Altarawneh, H. (2012). The state of using SMS-Based e-Government Services: Case Study in Jordan. *International Journal of Advanced Networking & Applications*, Vol. 4(3), pp. 1591-1600.
- [19] Mpinganjira, M. & Mbango, P. (2013). Profiling non-users of e-government services: in quest of e-government promotion strategies. *Journal Of Global Business & Technology*, Vol. 9(2), pp. 37-46.
- [20] Irani, Z., Weerakkody, V., Kamal, M., Hindi, N., Osman, I. H., Anouze, A., & ... Al-Ayoubi, B. (2012). An analysis of methodologies utilised in e-government research A user satisfaction perspective. *Journal Of Enterprise Information Management*, Vol. 25(3), pp. 298-313.
- [21] Venkatesh, V., Chan, F. Y. & Thong, J. L. (2012). Designing e-government services: Key service attributes and citizens' preference structures. *Journal Of Operations Management*, Vol. 30(1/2), pp. 116-133.
- [22] Khasawneh, R. & Abu-Shanab, E. (2013). E-Government and Social Media Sites: The Role and Impact. *World Journal of Computer Application and Technology*, Vol. 1(1), July 2013, pp. 10-17.
- [23] Criado, J. I., Sandoval-Almazan, R. & Gil-Garcia, J. R. (2013). Government innovation through social media. *Government Information Quarterly*, Vol. 30(4), pp. 319-326.
- [24] Barnard-Wills, D. (2013). Security, privacy and surveillance in European policy documents. *International Data Privacy Law*, Vol. 3(3), pp. 170-180.
- [25] Alhomod, S. & Shafi, M. M. (2013). A Study on Implementation of Privacy Policy in Educational Sector Websites in Saudi Arabia. *Global Journal of Computer Science and Technology*, Vol. 13(1), pp. 22-26.
- [26] Wu, K., Huang, S., Yen, D. C. & Popova, I. (2012). The effect of online privacy policy on consumer privacy concern and trust. *Computers In Human Behavior*, Vol. 28(3), pp. 889-897.
- [27] Stanaland, A. S. & Lwin, M. O. (2013). ONLINE PRIVACY PRACTICES: ADVANCES IN CHINA. *Journal Of International Business Research*, Vol. 12(2), pp. 33-46.
- [28] Alhomod, S. M. & Shafi, M. M. (2012). Privacy Policy in E Government Websites: A Case Study of Saudi Arabia. *Computer & Information Science*, Vol. 5(2), pp. 88-93.
- [29] Witte, D. S. (2014). Privacy Deleted: Is It Too Late To Protect Our Privacy Online?. *Journal Of Internet Law*, Vol. 18(1), pp. 1-28.
- [30] Allison, D., Capretz, M. A., ElYamany, H. & Wang, S. (2012). Privacy Protection Framework with Defined Policies for Service-Oriented Architecture. *Journal of Software Engineering and Applications*, Vol. 2012(5), pp. 200-215.
- [31] OECD (2013). The OECD website, OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data updated in 2013, accessed on April 26,2014: <http://www.oecd.org/internet/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm>

E-Government Adoption in Jordan: The Influence of Age

Nebal Q. Al-Jamal^a

Emad A. Abu-Shanab^b

MIS Department, IT College, Yarmouk University, Irbid, Jordan,
nebalaljamal@gmail.com^a & *abushanab@yu.edu.jo*^b

Abstract. The use of Internet and new technologies are driving governments to adopt the concept of e-government and work hard to attract all categories of citizens to use their online services. This study will explore the influence of age as a predictor of technology acceptance utilizing the original technology acceptance model constructs (perceived ease of use and perceived usefulness) and their influence on intention to use the technology. Results indicated a significant prediction of intention to use e-government services by age, perceived ease of use, and perceived usefulness. This paper supported the original technology acceptance model, and the role of age as a predictor of technology adoption. Age was negatively associated with of intention to use. Conclusions and future work are depicted at the end.

Keywords: *E-government, technology acceptance model, perceived ease of use, perceived usefulness, intention to use, age, Jordan*

I. INTRODUCTION

Information and communication technology (ICT) and the Internet are the driving force behind many new e-applications adopted by businesses and governments in this era. Governments around the world try to enhance their performance in order to increase citizen's satisfaction, and achieve their objectives successfully. Governments achieve their objectives and strategies through the use of several channels to deal with citizens to provide better services, reduces cost and efforts, make citizens more convenient, and empower them to effectively participate in the democratic process.

To benefit from the e-government concept, governments should guarantee equal citizens' access and knowledge of new technology. Such concept is called the digital divide. It is important to offer similar services, and provide equal opportunities to all categories of people in a country. Access and knowledge divides are major determinants that support the success of all e-government dimensions, such as e-democracy, e-voting and e-participation.

The main research objective for this paper is to explore the influence of age on the intention to use e-government services utilizing the technology acceptance model. The following section will review the literature, followed by a description of the research method followed. The next section will describe data analysis and discussion, followed by conclusions and future research.

II. LITERATURE REVIEW

The main research question of this paper is *"Would age significantly influence Jordanians intentions to use e-government services in the context of the TAM? Based on that,*

this paper will explore the concept of e-government, the influence of age and the technology acceptance model (TAM).

E-Government.

Governments around the world strive to reach their citizens to provide the needed services and empower them to participate in the democratic process. The e-government project in Jordan was launched in 2001 to transform Jordan into a knowledge society, and in order to contribute in the social and economic development [1]

Literature in this area doesn't provide a standard definition for e-government, where each definition concentrates on certain aspects of E-Government dimension. The basic notion of e-government revolves around using ICT tools and the Internet to provide better services to citizens. Also, some researchers stated that E-Government is not just a web site but it might be a powerful tool for empowering citizens and enhancing their life [2]

Abu-Shanab, p. 16 [3] defines e-governments as "the use of ICT, the Internet, wireless, and mobile networks, and web 2.0 tools and social networks to be able to perform the following: first, setup public polices, and apply them in decent, transparent, and in a high degree of accountability; second, provide a better services to citizens through all electronic means available; third, improve government's performance and efficiency through the necessary change and reengineering efforts; and fourth, reach out for citizens to fully participate in the political and social reform in an effective participatory, consultative and empowerment process. Such process is for the purpose of reaching good connected governance and society development".

E-Government main purpose is to construct a digital environment to provide citizens with electronic services and information they need through ICT tools. For the success of e-government in a country, governments should motivate citizens to use and utilize e-government services. In addition, there is a critical need to increase citizens' awareness about the benefits of using online services available on e-government websites [4].

Several studies addressed the benefits associated with the establishment of e-government projects. Foley and Alfonso [5, 21] stated that the most important benefit of e-government is the greater efficiency through saving money, minimizing labor cost, and providing better benefits for employees and various types of stakeholders. Citizens are considered an important part in e-government, where e-government contribute into reducing travel cost and time, reducing and saving citizen's time through the quick response, improved, more reliable, and up to date information, improving citizens service, providing convenience through the availability of multiple access channels, and improving citizen's service.

Research reported also several challenges and difficulties facing the adoption of e-government such as managerial and socio-economic factors, the digital divide, legislative issues, public governance issues, institutional complexity, trust in government and technology, and Psychological factors [4, 6, 7].

Even though more than 90 services are offered electronically to citizens and businesses on the website, still many Jordanian citizens are unaware of the full e-government services available and how to use them. Current Jordan e-government services include some of the following: allowing people in Jordan to obtain security clearances, background check certificates, renew commercial and professional certification at the Ministry of Industry and Trade, renew business licenses', and inquire about traffic tickets [8].

The Technology Acceptance Model (TAM)

The Technology Acceptance Model (TAM) is an information systems theory developed by Davis [9] and adapted from the Theory of Reasoned Action (TRA). It is considered one of the most important and popular models to anticipate end users acceptance and use of new technology. The TAM provides the basis for predicting the level of use of new technology, where external factors influence internal attitudes, intentions, and beliefs, and thus influence use. In its original version, TAM included five elements; perceived usefulness (PU), perceived ease of use (PEOU), attitudes (AT), intentions to use (ITU), and actual use (U). In the area of e-government, and for the Jordanian context also, research utilized more the influence of PU and OEOU on ITU [10, 11].

TAM determinants (PU, PEOU & ITU) are defined by Davis [9] and refer to technology in general. We propose the

following definitions for the purpose of adapting the constructs to the construct of e-government:

Perceived Ease of Use: PEOU is *the extent to which citizens believe that using e-government website is not complicated, and allow citizens to get the information and services they need effortlessly*'.

Perceived Usefulness: PU is *the extent to which citizens believe that using e-government website would be useful for their transactions, saves their time, and enhances their life*.

Intention to Use e-government website: *The extent to which citizens are willing to use e-government website and utilize services provided through it.*

Age as a determinant of e-government adoption

Age is an important demographic variable that has a significant impact on behavioral intention and acceptance of technology [12]. Age was explored in the context of using and adopting online services provided by e-government [13, 14, 15]. The fast introduction and evolution of new technologies are expected to lead to a differential level of adoption and use of such technology. It is expected that older citizens will have less skill and motivation to acquire and use of new technology [16]. Younger citizens are more likely to visit e-government websites and utilize its services than senior citizens [17].

The growing interest in e-government puts a heavy responsibility on government agencies to attract citizens of all age categories into adopting and using online services provided on e-government website. Warkentin et al. [15] proclaimed that citizens must have the intention to use e-government website and utilize the services provided for them in order to consider e-government initiatives as successful. Similarly, Sharma, Shakya and Kharel [18] stated that people's acceptance of e-government is a key driver of e-government success. They asserted that citizens, of all categories, should use e-government websites for a long period of time where they find it easy to use and provide them with the substantial benefits.

Based on previous research, Renaud and Van Biljon [19] proposed a senior technology acceptance model (STAM), which included several components such as: user context, experimentation and exploration, perceived usefulness, ease of learning and use, intention to use, confirmed usefulness, and actual use. According to their conclusions, the acceptance or rejection of a technology is determined by the ease of learning and use of such technology. STAM is a useful model, where it is provides an explanation of why many senior people do not fully accept new technology.

According to Bavarsad and Mennatyan [11] findings, the greater the ability of a government website to provide online

services the more citizen's use those services. Moreover, public agencies and governments should encourage citizens to use and accept online services through upgrading e-service provision, make big efforts to train and enhance citizens' awareness of such service so that all categories of users feel secure and easy to access and use e-services.

Porter and Donthu [13] concluded that seniors, and low income and less educated citizens have low e-government use and access than younger, high income and more educated citizens. This is attributed to citizens' individual beliefs toward Internet in general, and the effect of perceived ease of use and perceived usefulness of technology.

As summary age is one of demographic factors that affect the intention of citizens to use e-government websites. The United Nations e-government reports indicated that major Internet users are citizens of age less than 35 years, who speak English, live in urban areas and have fair education and income levels [20].

III. RESEARCH METHODOLOGY

This paper extended the TAM model with age, where three independent variables are predicting ITU. The research model adopted the robust PU and PEOU from the original TAM, and added age as an independent variable also. The only dependent variable proposed is ITU, where the depiction of this research is shown in Figure 1. The following hypotheses are developed and tested in this research:

- H1: PEU will significantly impact citizens' intention to use e-government web site.*
- H2: PU will significantly impact citizens' intention to use e-government web site.*
- H3: Age will significantly impact citizens' intention to use e-government website.*

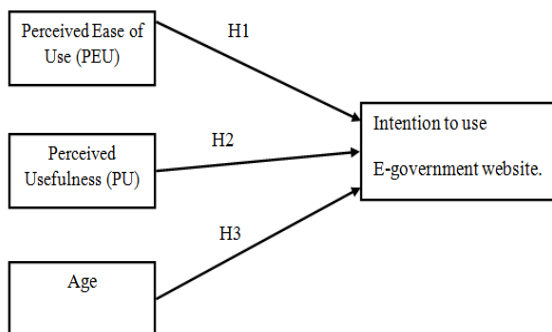


Fig. 1. TAM research model and hypothesis

In order to answer the research question and test the hypotheses, the authors developed a survey that includes items measuring the research constructs. The questions utilized a five-point Likert scale with 1 presenting “strongly disagree” and 5 presenting “strongly agree”. The survey developed based on the literature and improved after a pilot test. The survey was distributed in Arabic language due to the fact that respondents are Arab and Arabic language is the mother language of Jordanian citizens. The first part of the survey included demographic questions about the respondents' gender, age, and education.

The surveys were distributed by hand (paper 300 surveys) and online using Google documents (50 surveys). A stratified sampling process was used to guarantee a distribution of age categories, where citizens with age less than 25 years to above 46 years are targeted more. The total number of responses collected from both paper and Google surveys was 350. A visual inspection was conducted and removed 10 surveys with severe missing data to end up with a total number of usable surveys equal to 340. The demographic data of the sample is shown in Table 1.

TABLE I. DEMOGRAPHIC FREQUENCY STATISTIC

Gender	Count	%
Male	141	41.5%
Female	197	57.9%
Missing	2	0.6%
Total	340	100%
Age	Count	%
Less than 25	120	35.3%
25-35	57	16.8%
36-45	59	17.4%
More than 46	104	30.6%
Total	340	100%
Education	Count	%
High School & diploma	93	27.4%
Bachelor	196	57.6%
Master/PhD	51	15.0%
Total	340	100%

Data Analysis and Discussion

The means and standard deviations of all items measuring the TAM constructs are shown in Table 2. Also, the variable means are shown after each set of items, where the highest total variable mean is 3.43 (perceived ease of use), and the lowest is 3.23 (intention to use).

Regarding the highest and lowest mean of individual items, results indicated that item PU3 “Using e-government website allows me to access more government services.” yielded the highest mean (3.52), while PU2 “The results of using e-Government website are apparent to me.” yielded the lowest

mean value (3.20). Also, PEOU3 “Learning how to use e-government website to access government services is easy for me”, yielded the highest mean value (3.54), while PEU2 “My interaction with e-government website to access government services is clear.” yielded the lowest mean value (3.29).

Finally, item ITU2 “I predict to use e-government web site in the future.” Yielded the highest mean value (3.39), and ITU1 “I intend to use e-government web site continuously” yielded the lowest mean value (3.06).

The correlation matrix is used to determine if any two variables are associated to each other, and to check that independent variables are not extremely related. Table 3 depicts the Pearson Bivariate correlation matrix. Results indicate that age is negatively related to ITU, but not significantly related to PU and PEU. This means that as the age of citizens’ increase, their intention to use e-government website is decreased. On the other hand, a significant correlation exists between the three original TAM constructs.

TABLE II. DESCRIPTIVE STATISTIC FOR TAM ITEMS.

Item description	Mean	Std. Dev.
PEOU1: My interaction with e-government website to access government services is clear.	3.29	1.27
PEOU2: I find it easy to use e-government web site to find what I want.	3.49	1.08
PEOU3: Learning how to use e-government website to access government services is easy for me.	3.54	1.09
PEOU4: Overall, I find using e-Government website to access government services easy to use.	3.41	1.18
Perceived ease of use	3.43	
PU1: Using e-government web site enables me to access government services more quickly.	3.44	1.36
PU2: The results of using e-Government website are apparent to me.	3.20	1.27
PU3: Using e-government website allows me to access more government services.	3.52	1.23
PU3: Overall, I find e-government website useful for me to access government services.	3.31	1.41
Perceived usefulness	3.37	
ITU1: I intend to use e-government web site continuously.	3.06	1.30
ITU2: I predict to use e-government web site in the future.	3.39	1.19
ITU3: I plan to use e-government web site in the future.	3.22	1.36
Intention to use	3.23	

TABLE III. PEARSON BIVARIATE CORRELATION MATRIX.

Variable name	(ITU)	(PU)	(PEU)
Intention to use	1		
Perceived usefulness	.338**	1	
Perceived ease of use	.423**	.472**	1
Age	-.194**	-.082	-.074

** Correlation is significant at the 0.01 level (2-tailed).

The research also conducted multiple regression analysis to find the prediction level of ITU utilizing the three independent variables. Table 4 depicts the results of multiple regression analysis (the coefficient table), where the result of the ANOVA test yielded an R2 value equal to 381.812, with an F3,337 = 32.788, with a p value < 0.000. As shown in the table, the independent variables were statistically significant in predicting ITU. The standardized beta values of each variable are the following: PU = 0.169, with a p value less than 0.01; PEOU = 0.331, with a p value less than 0.001, and age = -0.155, with a p value less than 0.01. Such result indicates a full support of the original TAM and the research hypotheses proposed by this work.

TABLE IV. MULTIPLE REGRESSION ANALYSIS

	Unstd. Beta	Stand. Beta	t	Sig.
Constant	1.506		5.81	.000
PEU	.421	.331	6.06	.000
PU	.177	.169	3.09	.002
Age	-.132	-.155	-3.22	.001

IV. CONCLUSIONS AND FUTURE WORK

This paper supported the TAM findings, where perceived ease of use and perceived usefulness are significant predictors of the intentions to use e-government services. This result conforms to the findings of previous research [21]. Our

extension of TAM, age, yielded also significant prediction of ITU, but with a negative direction. Such results support our premise that older citizens will have lower intentions to use e-government services. Certain aspects of the TAM are emphasized more than others, but still all items used were perceived moderately by respondents (all means of items used were between 2.33 and 3.66, based on a 5 point Likert scale). Based on the comments of respondents and the results of this empirical test, the authors recommend that the success of e-government project is the responsibility of more than one party in community. The success of e-government projects depends on citizens adoption and attitudes, but still ministries that are responsible for providing such services are responsible too (like Ministry of ICT and other ministries that provide services through e-government website). The age divide might be the direct cause of this result, where governments need to build some effort to enhance the access level and the knowledge and skill needed to utilize e-government services through various ICT and the Internet. Government should pay more attention to the effect of age on the intention of Jordanians to use e-government services through the following propositions: holding workshops across Jordan to focus on the benefits and advantages of using e-government services especially to seniors, raise awareness to the importance of e-government projects, and conduct training programs to senior citizens on how to use technology.

This research is based on quantitative results, where a survey with forced choices to each item. Such issue limits the results of interaction of seniors with surveys and has some caveats. So future research should take into consideration this point and utilize qualitative methods such as interviews to get more and rich information from senior, where seniors in our culture like to talk more than read, and they express their opinions better through face to face setting. Also, as proposed by [22], other factors can contribute to the success of e-government that are related to environment and infrastructure.

REFERENCES

- [1] E-Gov Program, 2014. Internet: http://www.jordan.gov.jo/wps/portal/!ut/p/b1/04_SjzS3MLc0NTQ2NdSP0I_KSyzLTE8syczPS8wB8aPM4h2NHAMdPS2NDdzdvV0MPP1CA90cXZ0MDEwN9INT8_RzoxwVAXiDRZE/, [November 9, 2014].
- [2] B. AL-Rababah, and E. Abu-Shanab, "E-Government and Gender Digital Divide: The Case of Jordan," *International Journal of Electronic Business Management (IJEEM)*, vol. 8, no.1, pp. 1-8, 2010
- [3] E. Abu-Shanab, "Electronic Government: a tool for good governance and better service," Jordan: A book published by author (deposit number: 4441/9.2014, ISBN 978-9957-550-99-8), 2014, 257 pages.
- [4] Y. Dwivedi, V.Weerakkody, and M. Janssen, "Moving Towards Maturity: Challenges to Successful E-Government Implementation and Diffusion," *The DATA BASE for Advances in Information Systems*, vol. 42, no.4, pp.11-22, 2011.
- [5] P. Foley, and X. Alfonso, "E-government and the transformation agenda," *Public Administration* vol.87, no.2, pp. 371-396, 2009.
- [6] T. Nam, "Determining the type of e-government use," *Government Information Quarterly*, vol. 31, no.2. pp. 211-220, 2014.
- [7] E . Abu-Shanab, and A. Al-Azzam, "Trust Dimensions and the adoption of E-government in Jordan," *International Journal of Information Communication Technologies and Human Development*, vol. 4, no.1, pp.39-51, January-March 2012.
- [8] M. Ghazal. "Many Jordanians still unaware of e-gov't services – official. The Jordan Times." Internet: <http://jordantimes.com/many-jordanians-still-unaware-of-e-govt-services---official> . [November 8, 2014]
- [9] F. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," *MIS Quarterly*, vol. 13, no. 3, pp. 319-340, 1989.
- [10] E. Abu-Shanab, and A. Abu-Baker, "Evaluating Jordan's E-government Website: A Case Study," *Electronic Government: An International Journal*, vol. 8, no.4, pp. 271-289, 2011.
- [11] B. Bavarsad, and M. Mennatyan, "A Study of the Effects of Technology Acceptance Factors on Users' Satisfaction of E-Government Services," *World Applied Programming*, vol. 3, no.5, pp.190-199, 2013.
- [12] E. Abu-Shanab, J. Pearson, and A. Setterstrom, "Internet Banking and Customers' Acceptance in Jordan: The Unified Model's Perspective," *Communications of the Association for Information Systems (CAIS)*, vol. 26, Article 23, pp. 493-525, 2010.
- [13] C. E Porter, and N. Donthu, "Using the technology acceptance model to explain how attitudes determine Internet usage: The role of perceived access barriers and demographics," *Journal of Business Research*, vol.59, no.9, pp.999-1007, 2006.
- [14] W. R King, and J. He, "A meta-analysis of the technology acceptance model," *Information & Management*, vol.43, no.6, pp.740-755, 2006.

[15] M.Warkentin, D. Gefen, P.Pavlou, and G.Rose, "Encouraging Citizen Adoption of e-Government by Building Trust," *Electronic Markets*, vol.12, no.3, pp. 157-162, 2002.

[16] M. Redsell, and M. Nycyk, "Skilling seniors in computers: community training responses to the digital divide," *Working with Older People*, vol. 14, no. 2, pp. 38-42, June 2010.

[17] S. Hong, C. Lui, J. Hahn, J. Moon, and T.Kim, "How old are you really? Cognitive age in technology acceptance," *Decision Support Systems*, vol. 56, no. 2013, pp. 122-130, 2014.

[18] G.Sharma, S.Shakya, and P.Kharel, "Technology Acceptance Perspectives on User Satisfaction and Trust of E-Government Adoption," *Journal of Applied Sciences*, vol.14, no.9, pp. 860-872, 2014.

[19] K.Renaud, and J.Van Biljon, "Predicting technology acceptance and adoption by the elderly: a qualitative study," *In Proceedings of the 2008 annual research conference of the South African Institute of Computer Scientists and Information Technologists on IT research in developing countries: riding the wave of technology*, pp. 210-219, 2008, October

[20] UNDESA. United Nations e-Government Survey 2010: "Leveraging e-government at a time of financial and economic crisis." Published by the Department of Economic and Social Affairs, United Nations, New York, 2010.

[21] Abu-Shanab, E. Knight, M. & Refai, H. (2010). E-Voting Systems: A Tool for E-Democracy. *Management Research and Practice*, Vol. 2 (3), 2010, pp. 264-274.

[22] Abu-Shanab, E., Abu Al-Rub, S. & MdNor, K. (2010). Obstacles Facing the Adoption of E-government Services in Jordan. *Journal of E-Governance*, Vol. 33(1), 2010, pp. 35-47.

Big Issues for a Small Piece: RFID Ethical Issues

Mai Al-Sebae & Emad Abu-Shanab

MIS Department, IT College, Yarmouk University, Irbid, Jordan
maisebae@hotmail.com & abushanab@yu.edu.jo

Abstract. Radio frequency identifications RFID is becoming a widely used technology in many sectors. Despite of its benefits from tracking humans and things, it raises ethical concerns. This paper will explore research related to the ethical issues of RFID technology like privacy and security, coercion and green RFID issues. It is important to raise awareness related to the importance and seriousness of this technology, where it requires more legal control and support. This paper proposes a framework for future research to help researchers understand the issues embedded in such phenomenon and facilitate future research. Conclusions and future proposed research are depicted at the end.

Keywords: RFID, security and privacy, Coercion, IOT, Tag, Ethics, Green, proposed framework.

I. INTRODUCTION

Recently, Radio frequency identification (RFID) technology became one of the most influential and popular technologies that supports business objectives, where it started as a military tool for spying, and evolved to serve many fields like medical, social, and business. RFID technology uses radio waves to track individuals and objects and collect information about their behavior to help make important business decisions. Like any technology commonly used, its usage should be explored from economic, social, health and ethical views. In this paper, after defining RFID technology, its applications and its impact, we will focus on the ethical issues related to this technology, and its impact on people's acceptance of such technology

This paper will talk investigate the literature related to privacy and security of RFID, the coercive implant of this technology without the knowledge of the person carrying it, the justification and consequences of such use. Finally, the paper will explore green and environmental issues of RFID. A proposed framework for research is proposed to understand better this phenomenon. Conclusions and future research propositions are stated at the end.

II. LITERATURE REVIEW

A. What is RFID?

RFID is a technology for identifying people and objects automatically. RFID was at first introduced as a "sister" technology to replace barcode system for identifying items [1][2]. It has attracted considerable attention in recent years. RFID not only replaces the traditional barcode technology, but also provides additional features and eliminates boundaries that limit the use of previous alternatives [3]. RFID consist of tags, software and reader [4]. RFID tag is a small, microchip devised to transmit wireless data [5]. RFID tag is often formulated by a small chip for storing data and an

integrated transmitter antenna [1]. Tags are identified wirelessly by the readers by using some protocol of identification executed by readers and tags [6]. These tags operate at different frequencies and for that there is a need to employ different transmission mechanisms with different read-ranges, capabilities and bandwidths to penetrate sight line barriers. Still most of its use is for proprietary systems with specific use cases [7].

The technologies proliferation of automatic data acquisition, such as sensor and RFID technologies, aimed at improving the integrated environmental information systems (IEISs), decision support systems (DSSs), and environmental management. Such active and enduring topic for the scientists, engineers, and public administrators, involves broad issues beyond the use of many technologies [8]. Internet of things (IOT) is actuators and sensors mixed with the environment around us seamlessly, and shares information across platforms to develop a common picture for operating (COP) [9].

There are two types of RFID tags: active tag, which has embedded power source within it [1]. The second is passive tags, which doesn't have internal energy source (such as a battery). Passive tags get all the needed energy for functioning from an electromagnetic radiation transmitted by a reader. Communication between tags and reader is based on the "backscattering" principle in which the reader transfers energy to activate the tag, and then the tag replay by backscattering its identification data to the reader [10].

RFID technology has many advantages, such as improving real-time information traceability and visibility, quick reading, long distance recognition, and free obstacles[11][12]. Despite all RFID advantages there are some limitations and resistance from some parties especially organizations concerned with health, privacy, social and ethical issues. Cultural and religious factors may affect technology acceptability when related to RFID chips implanting in people for identification [13]. Helen Duce, director of the

RFID Centre at Cambridge University in England, says “We have a clear vision to create a world where every object from jumbo jets to sewing needles is linked to the Internet” [14].

B. RFID Applications

RFID is used by a variety of applications, ranging from supply chain management (SCM) to library systems [6]. Retailers convinced that applications like RFID can improve customer services, allowing for real-time inventory management, reducing customer checkout time, and providing information for customized shopping during and after sale [15]. RFID chips are implanted in animals to track livestock, study wildlife behavior and locate missing pets. Amusement parks utilize technology to find lost children, and schools manage students’ absenteeism. RFID also help locate, identify and monitor dangerous criminals, and also identify personnel military [4].

The success in animals-related applications has led to implanting RFID tags in humans. This might produce more dramatic outcomes than other application. However, its ethical issues become debatable when deciding the range of use/abuse because of RFID invasive nature [16]. RFID can be used in management of material, machinery, and men (M3). Management of material includes logistic and SCM, Inventory management, quality assurance, and waste management. Management of machinery consists of tracking machines and tools, machine operation and records, and machine maintenance records. Management of men includes access control, labor attendance records, and men safety [1].

From an ethical point of view, the most important issue is how such technology is used. Governments have considered RFID to track and control released parolees, convicts, and foreign visitors [16]. In 2003, Wal-Mart enforced the adoption of RFID technology for its suppliers. RFID can be applied in inventory monitoring at retail-stores and it can be helpful for products replenishment on real time basis [17]. RFID applications are most likely to increase healthcare environment reliability [18], however, there are obvious privacy and security concerns related to storing medical and personal data. On the other hand, securing authentication of healthcare system/environment is still a challenge as it touches on issues of confidentiality, unforgeability, scalability, and location privacy issues [19].

RFID systems can be used in hospitals to verify patient’s identity during medical procedures, locate equipment and collect data from staff workflow. RFID systems allow for the electronic tagging of inventory, assets, personnel, and patients, but there is a little empirical evidence on how to effectively implement such systems. RFID systems are not easily adapt to hospital settings because of its complicated infrastructure (in terms of equipment, space, personnel, and patients). Hospitals implementing RFID systems tend to experience two constraint types: 1) the technological system maladaptation to hospital settings, and 2) the organizational challenges for hospitals when utilizing the system [20].

Manufacturers are already using RFID technology in products that could putrefy during transportation due to extreme temperatures. The RFID reader will respond with data that indicates the state of product as well as its ID, when it interrogates the tag [3]. Vehicle Traffic Congestion Estimation (VTCE) is another application of RFID, where a reader reads the tags installed on vehicles and transfer the necessary data to a database in a Central Computer System (CCS). CCS utilizes the data to locate the traffic congestion by following a specific procedure. However, it needs a high implementation cost compared to the device, maintenance and installation cost [21].

C. Ethical Issues Of RFID

It’s very important to judge any technology in terms of its ethical perspective because the compatibility and harmony of technology with morality (ethics) has a significant impact on its acceptance by the individuals and organizations at large. Three types of ethics are identified: computer ethics, information ethics, and cyber ethics [22]. This applies to the all ICT applications. Ethics aims to provide means to help discern what humans should do and how they should behave. It guides us to what is good or bad behavior, and deals with behaviors or actions rather than thoughts or feelings. It is also realized that what is an acceptable behavior in one culture might not be ethical in another [23].

Investigative panels in USA (including engineers and technicians developing RFID, military and commercial interests, and experts in ethics) should examine the directions of RFID and if the technology should be allowed to do that or not [16]. The following sections will examine three types of ethics that are related to RFID based on what has been reported in previous studies and they are: privacy and security, green RFID and Coercion of use.

i. Security and privacy

Research about ethical issues focused more on Internet security, like data theft and personal information hacking, and especially in business domains [23]. [24] Asserted that security is the protection of information systems holdings and controlling information access; he added that privacy means respecting individual’s right to have information with an appropriate protection level. Despite the advantages of RFID, its application may challenge the privacy and security of organizations and individuals. However, the use of RFID systems raises new privacy and security issues. For RFID systems to become widely accepted by industry and end-users, security and privacy preserving authentication protocols are required [6].

The RFID introduction adds a new dimension to debates over consumer’s privacy by allowing tracking of products after the point of sale. The issues at pin depend on two factors: how the item is considered to be personal, and the item mobility. It can be argued that ethical concerns arise more from tagging a library book than tagging a soup can. Greater privacy risk comes from embedding a tag in eyeglasses worn by a consumer continuously, than from embedding a tag in a

"sofa" that remains in place. However, RFID does not break privacy any more than bar codes and credit card use, unless hackers have access to readers and to the associated databases [4].

Several threats to RFID systems are reported like: Eavesdropping which means listening in secret to communication between reader and tag by an illegal user, which is resolved by producing changing values that don't allow attackers to access significant data if they acquired it [11]. Secondly, traffic analysis to intercepting messages to deduce valuable data from patterns in communication between the tag and reader, which can be solved by adding random number in the tag and reader communication information. Thirdly, replay attack means valid data transmission is fraudulently delayed or repeated, it's carried out by an adversary who retransmits the data and intercepts or by the originator. Finally, one of the most serious privacy problems of RFID systems is tracking attack by attacker who can track the user's location information and risk users privacy, which can be solved using random number or timestamp.

Another solution to protect privacy is a privacy-preserving authentication "DPM" protocol introduced by [6] for RFID tags; DPM introduces a repeated identification technique. The tag sends several randomized secret hash keys, and the tags are identified by the reader by eliminating entries that do not match the hash result in its database successively. Despite the advantage of the DPM-protocol by reducing the complexity during the authentication on the reader-side, a major DPM-protocol drawback, and many others, is in the requirement to evaluate a strong encryption hash function.

The RFID technology users can be partitioned into individual, business and government. All of these classes of users can use RFID in a different way that proved beneficial to the generic population. Private sector enterprises can generate revenue from the technology. Also, consumers can benefit from technology by reduced prices and enhanced quality of service. Both of them can make tradeoffs between convenience and privacy when it is the time for opt-out options [4]. RFID privacy concerns with misbehaving readers and the problem of harvesting information from tags behaving well [11].

Issues of data security continue to be a challenge to this industry's growth. Retailers wish to use RFID to provide improved services, while customers may be afraid of entering environments and interested in transactions that could compromise their privacy. Customers' trust is the most important factor for retailers who aim to establish an RFID item-level retail store [15].

[7] Conducted a study to distinguish between three approaches to address consumers' concerns of privacy: First, to kill RFID tags at store exits (business-controlled). Second, if they want to initiate reader communication, then lock tags and have user unlock them (user controlled model). Third, to let the network access users' RFID tags while privacy protocol adhering (network model). The author concluded that the majority of consumers want to kill RFID chips at

store exits rather than using any presented complex technical solutions. In addition, the desire to kill RFID tags is not caused by the fact that consumers do not recognize the value or benefits of RFID services but because of its seamless and ease of use of such approach. Despite of the service value which can be realized by RFID, customers are willing to forget these advantages in order to preserve their privacy.

RFID must attain some goals regarding privacy and security. The following are reported in the literature: Tags must not compromise holder's privacy, information should not be available to unauthorized readers or be possible to build long term tracking connection between tags and holders, holders should be able to find and disable any tags they load to prevent tracking, the output of publicly available tag should be easily or randomized modifiable to avoid the long-term connection between holders and tags, contents of private tag must be protected, both readers and tag should trust each other, spoofing either party should be hard, provide a control access mechanism and a mutual authentication between readers and tags also provides a measure of trust, and finally, concerns of session hijacking and replay attacks need to be resolved [25].

ii. Coercion

Receiving an RFID tag is purely a matter of consumer choice, and raises few serious ethical issues. The most important ethical issue is the possibility that the chips might be implied or implanted under real coercion, with the deep aversion or at least unease with many individuals [13]. Three concerns are raised: First, health risks to patients, like emitting radio waves that could cause migrating throughout one's body, tumors, or requiring surgery to be removed. Second, privacy; explained in detail earlier. The third is coercion, which means that patients (subjects or customers) are chipped without knowing that [5]. For example people with dementia are at risk for this reason as they fall within the most commonly marginalized groups in society; they are older, generally female and have concerns of mental health [26].

[27] Concluded through a study about tracking dementia (Alzheimer's disease) through a transmitter that some patients were happy to wear the transmitter, but later rejected it because they did not want to be kept on a lead. A more relevant ethical justification was that the RFID device had the ability to reduce the time spent to find lost patients, and lessening the chances that would cause an accident. RFID also can identify patients and improve the accuracy and streamlined delivery of health care, but they may also introduce new ethical, medical and social, risks. In general concerns about implants have been largely theoretical and concentrate on the devices safety, patients' records privacy, and coercion to consent to the devices implantation [5].

Decisions about using RFID need to be made by a broader group of stakeholders than engineers and companies involved in the field. A commitment must be made to enclose the technology to people who choose freely to carry it and protect others from implied or coercive implantation [13]. But in cases when RFID could aid in finding missing persons,

kidnaped victims, lost hikers, lost high-profile personnel and children there's a debate if people are chipped in coercion or by their choice [16]. Tracking such categories is necessary to protect them or protect others from them.

iii. Green RFID

Green IT is the study and practice of manufacturing, designing, and using computers, monitors, servers, storage devices, printers, networking, and communication systems effectively and efficiently with the least impact on environment. It includes environmental sustainability, energy efficiency economics, and total cost of ownership. It incorporates the disposal and recycling cost of technology. Green IT is also about IT application to create energy-efficient, environmentally sustainable business practices and processes. IT can support, leverage and assist other environmental initiatives and create green awareness also[28]. RFID could indeed be green by improving environmentally responsible practices related to IT. Organization should not focus only on adding "economic value" (such as preventing its spoilage and accurately tracking a perishable item), but save operations' energy ranging from growing, harvesting, packaging and refrigeration [29].

A study of 13 cases that used green RFID (such as Wal-Mart, Nestlé Italy, Rewards for Recycling, Recycle Ban, Concept2Solution, Smart Vareflyt, Truck Tag, DHL Smart Truck, Strawberry, Indisputable Key, The City of London School for Girls, Multi Life Cycle Center and Promise) employed the (MECI) framework which is a short cut for green categories (Motivation, Execution, Challenges and Impacts). Motivation refers to the reasons that drive the green RFID projects adoption such as financial, operational and strategic reasons. Execution; refers to processes and Strategy that get executed during the project such as Unit level, Project progress and Partnerships. Challenges; refer to the Difficulties (technological and informational) faced during project implementation. Impacts; refer to unexpected and anticipated consequences from project implementation like environmental sustainability, business value and Social responsibility. Results indicated that the RFID challenge is to sacrifice profit for environment benefit and vice versa.

In order to consider any development sustainable, it should meet the present needs without compromising the future generation's ability to meet their needs[30]. Applications related to the development of RFID antenna were ignored while facing the interconnected issues of the eco-friendly and economic field tag comprising substrate [31].

[28] add that there are two waves of green IT: 1) Internally focused on IT products reengineering and processes to meet compliance requirements and improving energy efficiency. 2) Externally focused on sustainability-based IT innovations, business transformation, and enterprise-wide sustainability. The alignment between ICT processes and practices with the three core sustainability principles (reduce, recycle, and reuse) and innovatively using ICT in business processes is the key to leverage benefits of the enterprise-wide sustainability.

[10]talked about antenna RFID as a "green" electronics solutions, which is robustness, flexibility and eco-friendliness, outstandingly the proposed read range antennas make them an absolute choice for industrial far field. "green RFID" is challenging because the Increase in complexities of operational processes and the possibility of sacrifice on economic benefits have limited the diffusion of RFID technologies as green[32].

In an experiment on the city of Grand Rapids for using RFID for green purposes, RFID tags allow for signing reject cart recycling and its related data of a specific customer and location. This provides a detailed ownership history, cart repairs, location, and allows monitoring city's cart asset during its useful life. They have now approximately 77 thousand carts on the street. They're able to produce specific participation rates to recycle refuse customers, something they weren't able to do with the reject bags; they were able to determine incinerator landfill tonnage diversions. They were able to measure efficiencies rout and allow for operational decisions to be quickly made. Finally, it maximized revenues from operating from recycling collections and refuse [33].

The literature reports some applications of green RFID projects; in Paris, more than one hundred thousand trees have an RFID tag that allows for an easier monitoring by municipality officials. Nestlé uses control temperature mechanisms to improve quality, reduce spoilage, and lower energy costs. Wall Mart Stores are working with men's jeans suppliers using RFID tags to be able to track these items to optimize its inventory management and indirectly reduce CO2 emissions [34]. There are two approaches which may improve RFID support of green initiatives and sustainable development projects through the commitment to greening policies in RFID design or by contributing to some of the greening projects in general.

III. CONCLUSION

The main benefit of RFID is to identify people and things. It can largely enhance real-time information for decision making purposes. RFID is mostly used in SCM, health care, traffic, and libraries. Researchers interested in the ethical side of technology are focusing more on privacy, security, coercion, and green issues. Privacy and security are attracting more research, while coercion is affecting the success or failure of such technology use. RFID raises a debate about the purpose and justification of this technology use without telling individuals. There is also great interest in green projects in line with institutions' care to highlight their interest in social responsibility and environment-friendly use of technology, where RFID provides great contribution for greening projects.

As we mentioned that RFID applications are becoming multi-purpose in helping organization and governments to track people, animals and objects. Each one of these categories has many uses and applicable examples as we summarized. The significance and sensitivity of privacy and security issues are increasingly attracting attention more than tracking animals

and objects (unless these things are closely related to individuals). Similarly, the coercion issue is less significant when related to tracking objects and animals compared to humans.

This paper proposes a framework that map all the constituents related to the application of RFID. Such mapping guide future research and focus efforts towards a more comprehensive coverage of all issues related to the ethical issues of RFID. The frame work is built around three major dimensions: people, animals, and objects. People will be influenced by the security, privacy, coercion, and green IT issues. Humans are the most sophisticated dimension as they guard for future generations, and the environmental issues related to RFID. Animals are more related to environmental issues. Finally, objects are related to humans if they were implanted and used by humans, so some issues are common.

Research in RFID area needs to weigh the benefits of RFID vs. the cost of green IT issues and the risks of privacy protection. Such tradeoff is important to understand that the benefits realized from RFID need to be balanced against its risks and challenges. Keeping in mind the efforts by organizations in regard to social responsibility, it is important to keep revenue from RFID at acceptable levels. Finally, the RFID design needs to guard for the cycle of recycle, reuse and reduce.

This framework setup the stage for future research which can focus on areas depicted at the heart of the diagram. Also, technology adoption issues and trust are important to guarantee its success. Once researchers understood all aspects related to RFID implementation, they can better serve organizations employing it and future generations.

Fig. 1 How ethics is related to RFID

REFERENCES

[1] Lu, W., Huang, G. Q., & Li, H. (2011). Scenarios for applying RFID technology in construction project management. *Automation in Construction*, vol.20(2), pp.101-106.

[2]Bunduchi, R., Weisshaar, C., & Smart, A. U. (2011). Mapping the benefits and costs associated with process innovation: The case of RFID adoption. *Technovation*, vol.31(9), pp. 505-521.

[3]Want, R. (2004). Enabling ubiquitous sensing with RFID. *Computer*, vol. 37(4),pp. 84-86.

[4]Glasser, D. J., Goodman, K. W., &Einspruch, N. G. (2007). Chips, tags and scanners: Ethical challenges for radio frequency identification. *Ethics and Information Technology*, vol.9(2), pp.101-109.

[5]Monahan, T., & Fisher, J. A. (2010). Implanting inequality: Empirical evidence of social and ethical risks of implantable radio-frequency identification (RFID) devices. *International journal of technology assessment in health care*, vol. 26(04), pp.370-376.

[6]Blass, E. O., Kurmus, A., Molva, R., Noubir, G., &Shikfa, A. (2011). The F_f-Family of protocols for RFID-privacy and authentication. *Dependable and Secure Computing, IEEE Transactions on*, vol. 8(3), pp. 466-480.

[7]Spiekermann, S. (2009). RFID and privacy: what consumers really want and fear. *Personal and Ubiquitous Computing(Springer)*, vol.13(6), pp.423-434.

[8]Fang, S., Da Xu, L., Zhu, Y., Ahati, J., Pei, H., Yan, J., & Liu, Z. (2014). An Integrated System for Regional Environmental Monitoring and Management Based on Internet of Things. *IEEE Trans. Industrial Informatics*, vol.10(2), pp.1596-1605.

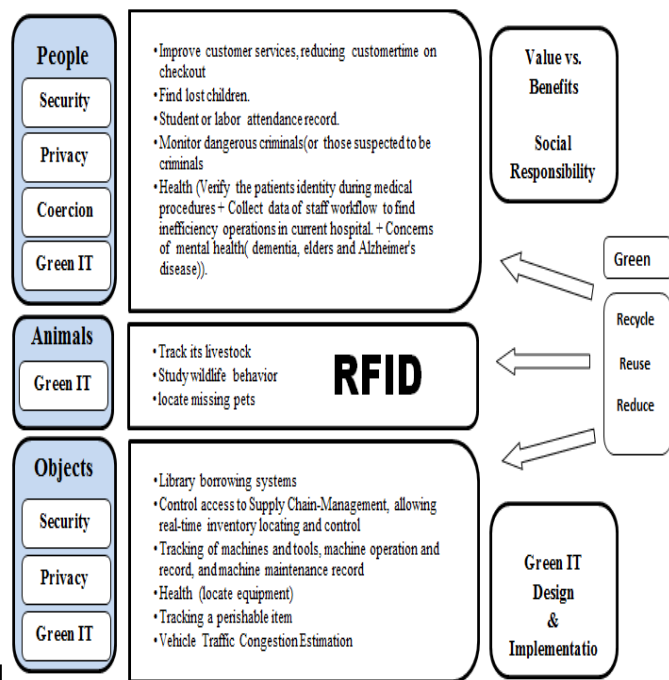
[9]Gubbi, J., Buyya, R., Marusic, S., &Palaniswami, M. (2013). Internet of Things (IoT): A vision, architectural elements, and future directions. *Future Generation Computer Systems*, vol.29(7), pp.1645-1660.

[10]Amin, Y., Chen, Q., Zheng, L. R., &Tenhunen, H. (2012). Development and Analysis of Flexible UHF RFID Antennas for`Green. *Progress In Electromagnetics Research*, vol. 130, pp.1-15.

[11]Wang, M., & Pan, J. (2014). Authentication Test-Based the RFID Authentication Protocol with Security Analysis. *Sensors & Transducers*, vol.176(8), pp.1726-5479.

[12]Abu-Shanab, E. (2013). RFID Utilization in Libraries: The Benefits, Challenges and Future Directions. *First International conference "Libraries and Information Centers in a Changing Digital Environment,"* Amman, Jordan, October 29-30, 2013, pp. 1-17.

[13]Foster, K. R., & Jaeger, J. (2008). Ethical implications of implantable radiofrequency identification (RFID) tags in humans. *The American Journal of Bioethics*, vol. 8(8), pp. 44-48.



- [14]Smith, R.,E.(2013). New Cause for Concern: ‘Internet of Things. *Privacy Journal*, vol.39, pp.3- 5.
- [15]Yeh, H. (2013). Effects of RFID in Retailing on Customer Trust. *Journal of Service Science and Management*, vol. 6, pp.143-150.
- [16]Labay, V., & Anderson, A. M. (2006). Ethical considerations and proposed guidelines for the use of radio frequency identification: Especially concerning its use for promoting public safety and national security. *Science and engineering ethics*, vol. 12(2), pp.265-272.
- [17]Modgil, S., Patyal, V. S., &Agrawal, T.(2012). Technology Evolution in Supply Chain Management: BARCODE to RFID. *CPMR-IJT International Journal of Technology*, Vol. 2, (1), p.p27-32.
- [18]Awawdeh, N. & Abu-Shanab, E. (2009). RFID in Healthcare Industry: A Strategic Framework for Implementation. A conceptual paper presented in the *Proceedings of the 4th International Conference on Information Technology (ICIT 2009)*, Amman, Jordan, pp. 1-18.
- [19]Wu, Z. Y., Chen, L., & Wu, J. C. (2013). A reliable RFID mutual authentication scheme for healthcare environments. *Journal of medical systems*, vol.37(2), pp.1-9.
- [20]Fisher, J. A., & Monahan, T. (2008). Tracking the social dimensions of RFID systems in hospitals. *International journal of medical informatics*, vol.77(3), pp. 176-183.
- [21]Al-Naima, F. M., &Hamd, H. A. (2012). Vehicle Traffic Congestion Estimation Based on RFID. *International Journal of Engineering Business Management*, vol.4(2),pp .1-8.
- [22]Ramadhan, A., Sensuse, D. I., &Arymurthy, A. M. (2011). e-Government Ethics: a Synergy of Computer Ethics, Information Ethics, and Cyber Ethics. *International Journal of Advanced Computer Science & Applications*, vol.2(8), pp.82-86.
- [23]Salman, A., Saad, S., Ali, M., &Shahizan, N. (2013). Dealing with Ethical Issues among Internet Users: Do We Need Legal Enforcement. *Asian Social Science*, Vol.9(8), pp.3-8.
- [24]Fang, Z. (2002). E-government in digital era: concept, practice, and development. *International journal of the Computer, the Internet and management*, vol.10(2),pp. 1-22.
- [25]Sarma, E. S., Weis, S. A., & Engels, D. W. (2002). White paper: RFID systems, security & privacy implications. Cambridge, MA: Massachusetts Institute of Technology, *AUTO-ID Center*.p.p1-16.
- [26]Plastow, N. A. (2006). Is Big Brother watching you? Responding to tagging and tracking in dementia care. *The British Journal of Occupational Therapy*, vol. 69(11), pp.525-527.
- [27]McShane, R., Gedling, K., Kenward, B., Kenward, R., Hope, T., & Jacoby, R. (1998). The feasibility of electronic tracking devices in dementia: a telephone survey and case series. *International journal of geriatric psychiatry*, vol.13(8), pp.556-563.
- [28]Laplante, P. A., &Murugesan, S. (2011). IT for a Greener Planet. *IT Professional, I E E E Comp u t e r S o c i e t y*, vol. 13(1), P.P 16-18.
- [29]Bose, I., & Yan, S. (2011). The green potential of RFID projects: A case-based analysis. *IT professional, I E E E Comp u t e r S o c i e t y*, vol.13(1),pp. 41-47.
- [30]Corvalan, C., Hales, S., & McMichael, A. J. (2005). Ecosystems and human well-being: health synthesis. World health organization.
- [31]Amin, Y., Chen, Q., Tenhunen, H., &Zheng, L. R. (2012). Performance-optimized quadrate bowtie RFID antennas for cost-effective and eco-friendly industrial applications. *Progress In Electromagnetics Research*, vol. 126, pp. 49-64.
- [32]Yan, S., & Bose, I. Macau,(November 30 - December 4, 2009) RFID AS GREEN IT: LESSONS LEARNED FROM CASE STUDIES. *The 9th International Conference on Electronic Business.*,
- [33]Angeles, R.(2013). Using the Technology-Organization-Environment Framework and Zuboff’s Concepts for Understanding Environmental Sustainability and RFID: Two Case Studies. *International Scholarly and Scientific Research & Innovation*, Vol.7 (11), pp,1605-1614.
- [34]Duroc, Y., &Kaddour, D. (2012). RFID potential impacts and future evolution for green projects. *Energy Procedia(Elsevier)*, vol.18, pp. 91-98.

The Effect of Using Social Media in Governments: Framework of Communication Success

Dareen A.Mishaal¹ & Emad Abu-Shanab²

MIS Department., IT College
Yarmouk University
Irbid, Jordan

damshal05@aabu.edu.jo¹
abushanab@yu.edu.jo²

Abstract— The vast emergence of social media with its characteristics and benefits opened doors for individuals and groups to connect utilizing this revolution. Social media became a huge virtual community, with highly interactive and collaborative environment among its members. Governments realized that more and more of their citizens are present over social networks, not over governments' websites. This paper takes Facebook as one of social media applications and builds a framework for measuring communication success over social networks. The model proposes that transparency, participation, collaboration and comfort will lead to communication success. Also, the model assumes that the posted topic will influence communication success. Finally, we propose indicators and metrics to measure factors proposed in the model.

Keywords— *government; social media; Transparency; participation; collaboration; comfortable; posted topic; communication success framework.*

I. INTRODUCTION

The ubiquity of the Internet is becoming an important phenomenon that changed the world. Internet influenced every aspect of private and public lives, and changed the nature of service towards a click and mortar instead of brick and mortar nature. Governments recognized the importance of Internet, and started to provide their services electronically over e-government websites. This initiative was to increase the performance of government services provided to different stakeholders (citizens, businesses and government itself) [14]. As the Internet evolved, the government also evolved in the same direction and witnessed a move from e-government services to social government; i.e. governments provide their services over social media such as Facebook, twitter, LinkedIn, and Flickr [13]. Social media is defined as "a group of Internet-based applications built on the ideological and technological foundations of Web 2.0 that allow for the creation and exchange of User Generated Content" [12]. Social media includes Facebook, Google plus, twitter, blogs, wikis and YouTube, where all of them are built based on web2.0 technology. Social networks are defined as "a networked platform, spanning all connected devices that encourage collaboration in terms of the creation, organization, linking and sharing of content" [23]. Web2.0 technologies have a set of capabilities grouped in the word **SLATES**, which enable the organization to search for employees to determine resources effectively and enable the

organization to link employees with customers, and authoring by enabling the employees and customers to collaborating in creating and sharing contents. Tagging also enables people to organize and filter the content, where these technologies enable the extensions by enabling the share of complex multi-media content and signals for the deployment of the changes over the content [8].

The main objective of this paper is to explore the benefit of social media for governments. For this purpose, we explored the available literature related to same topic and organized the work as follows: the first section of this paper will report the literature related to the social media and its contribution to e-government objectives The following section will describe the methodology, followed by conclusions and future work.

II. LITERATURE REVIEW

The main research question of this paper is *what are the factors that make governments successful in communicating with citizens and businesses using Facebook?* Based on this question this paper proposed a framework for government's communication success utilizing Facebook, and also proposed a set of measures and metrics for the model to be tested for different governmental Facebook pages.

a. Deficiencies of Traditional Government Communication

Hofmann, Beverungen, Räckers and Becker [11] argued that governments should provide information to their stakeholders

in a way to communicate with them. But many governments have problems in their communication due to low budget and because they put communication with stakeholders as a low priority. Also, governments use the traditional methods of communication such as newspapers, radio and television; this one-way communication reflects the low feedback of stakeholders to communicate with government, which leads to low participation from stakeholder's side.

Governments in the Arab world suffer from many challenges or deficiencies, some of them are internal and others are global or regional. Governments have multiple deficiencies which include the low trust in government, limited accountability, lack of transparency and low quality of service related to deficiencies in accessibility to the different services [6] [11]. Research indicated that Arab governments have deficiencies in their democracy and sometimes they lack such phenomenon at all. However, Islam as a religion adapted democracy through "Shura" (consultation) and ignored the autocratic style of governance [15] and Shura in the first Islamic era is compatible to the democracy in modern political life.

b. Social Media Detentions

The concept of social media nowadays became one of the hottest topics that attract consultants, experts, and researchers, where they can gain benefits from using it in their fields [12]. Social media changed the way developers and users are using the web; it shifted the domain from a self-content generation to a group-content generation, based on the level of participation and collaboration between them.

Margo asserted that social media takes its importance from its characteristics that include: participation, collaboration, empowerment, and time [18]. It enables people to participate and share the content between each other. Also, it enables users to collaborate by creating their own community to achieve their goal. Social media empowers users to participate with their ideas and opinions by enabling a platform to discuss their idea and opinions which in other words promote democratization. The time character in social media enables users to publish their generated content and publicize it to be seen by their friends at the same time of publishing, which enables others to participate instantly with their opinions and comments.

Snead classified social media into internal-based and external-based technology. The internal-based technology runs over the agency's server such as blogs and wikis updated based on the agency's needs. The external-based technology runs on servers of a third party, where the agency has no control over it, such as Facebook and twitter [25].

Social media is a wide umbrella that includes many types based on their purpose and functions [6] and in the following fashion: 1) directories 2) communication channels 3) communities and rating sites, and 4) archiving and sharing sites. LinkedIn is an example of directories; its function is to type the resume listing and rating by clients and colleagues. Twitter and blogs are examples of communication channel,

which is used for publishing information and text in real time. Communities and rating sites such as Facebook are used for interaction in a closed site, and it is less formal. Finally, archiving and sharing sites such as YouTube, are used for saving, sharing and redistribution of documents, videos, and slides. Facebook, our focus in this paper, is a social network launched in the year 2004 for the students of Harvard University. Facebook enables users to interact with other users by changing their status, writing on the walls of other users, and sending personal messages. Facebook enables users to create their groups, join other groups, or like other pages. Also, it enables users to upload their photos and videos, and search for contacts and content [25] [26].

c. Importance of Social Media in Governments

The impact of social media on government has taken place from the year 2011, especially through the Arab spring, and it was responsible for changing the relationship between governments and citizens [27]. Social media enables the democracy to be shared internationally, enables the citizens to express their opinions, and the governments to get feedback from their citizens. Using social media channel between the government and its stakeholders will open the dialogues between the government and its citizens, which started by disseminating information to public via social media and by listening to citizen's opinions and feedback on such information. The popularity of social media and its characteristics enables citizens or any stakeholder to get more information that makes them more acquainted by governments' activities [20] [22] [23].

Social media when activated very well will affect government's performance positively in tracking the opinions and mood of public, and instead of using traditional ways of collecting responses from public, governments can utilize such information for more effective decision making [26]. Other researchers indicated that most managers started to use the reports and insights generated from social network sites to prepare their reports instead of developing their own metrics or indicators to study the behavior of citizens and to get citizens' opinions by using social network, which is also more cost effective [20] [25]. Using social media in e-government will enable citizens to access information provided by their governments over social media and thus reducing the effort for searching for needed information [7].

The major benefits of using social media in e-government can be summarized in three benefits [19] [20]: 1) Transparency, which is the release the information that stakeholders are always checking for. 2) Participation is to maintain citizens engaged with their governments, by allowing citizens to express their opinions, experiences and wisdoms. Governments can use a survey to get feedback from citizens before any decision is taken, then pass this information to different government agencies for actions to be taken. 3) Collaboration is the high level engagement between government and citizens, where citizens participate by

creating the content of government topics and the government use and follow the content generated by citizens to fulfill government mission [27]. This way more time is saved and thus cost and effort for governments [16].

Research indicated that transparency is significantly related to the advancement of e-government [5]. On the other hand, e-participation includes five levels where collaboration and empowerment are the highest levels of e-democracy [4].

d. Government and Social Media in Communication

Social media can be considered as an innovative technology for governments, which added to e-government extra benefits. Klischewski [15] argue that social media needs to maintain the relationship between citizens and their government; this needs self-discipline to effectively and consciously use such media in a free and open way. Also, it is important to shape the communication itself in a way to determine the information and collaboration requirements. Finally, it is important to shape the social media itself to guarantee the suitable use of such media in different political activities. Social media was adopted in U.S. federal government according to the guidance of Barak Obama to find ways to make the information open and public to all stakeholders. Such initiative included directions to increase transparency, participation and collaboration. The government disseminates the information to the public and thus information become available to all; this enables citizens to participate and collaborate with each other [21] [22].

It is realized from our experience in the Arab spring that citizens need a democracy because of the domination of autocratic leaders for long periods of time [15]. Citizens need to participate freely in public debates, need transparency and accountability to be emphasized more by governments. Al-Saggaf and Simmons claim that Arab countries are different from the western countries, where their citizens are engaged in political life and express their opinions in many ways [5]. The utilization of social media enhances the traditional way of communication between the government and their public. Because Arab governments have limited freedom for their citizens to express their dissatisfaction, the new evolvments in social media was the only way to communicate with governments. Social media opens a new political place for Arab citizens; it enables citizens to express their opinion in public debates and interact without any controls [14]. The adoption of social media in e-government has other determinants such as political context and culture [5]. The political context and culture will determine the nature of public participation.

e. Government-Social Media Based Models

Government-social media based models are frameworks produced by researchers to understand the importance and benefits of social media in e-government. The first models assumes that social media is a catalyst that transforms citizens,

government and data [8]. It discusses how social media interacts with these three dimensions. *Social Media-Based Citizen Engagement Model* focuses on social media used as a tool to enable users to express their opinions, emotions, behaviors and interactions. Governments can use such media to transform their citizen to participate in god governance and to enjoy democracy. *Social Media-Based Data Sharing Model* focuses on the data that a citizen generated using social media when he/she participated in a political topic. This data needs to be understood by citizens and governments. Also, it needs to be stored and processed to make it sensible and usable for making decisions, and to enable citizens to participate collaboratively with governments. *Social Media-Based Real-Time Collaborative Government Model* focuses on the idea that using social media in e-government starts to enhance the communication between governments and citizens to be nearly in real time.

Lee & Kwak [16] proposed a maturity model for open government that is based on public engagement. The model contains five levels, which is based on the benefits of social media that provide transparency, participation and collaboration by public engagement. The following is a description of each level:

- *Level 1: Initial Condition*, this level focuses on government podcasting information to public via government's official websites, with seldom use of social media or interactive tools.
- *Level 2: Data transparency*, actually is the first step of open government, where governments try to perform two important tasks, first is publish valuable and impact data, second is to publish data that is accurate, consistent and in a timely manner.
- *Level 3: Open participation*, this level focuses on enabling the public to participate and governments to take input from citizen's feedback, participation, discussion, and voting.
- *Level 4: Open Collaboration*, is the developed level of participation, where the government asks the citizen to play a role of co-creation, co-design for specific output. Here the task is more complex, and it looks like collective intelligent, so government may use shared document to engage the public in participation, asking the public to participate in designing application to government.
- *Level 5: Ubiquitous Engagement*, this level is built based on level 1 to level 4 by expanding the level to arrive to engagement status. In this level we can see public engagement become easier by using different accessing technologies such as smart phones, tablets, laptops and desktops.

f. Influencers of the Adoption of Social Media in e-Governments

The adoption of social media by governments is an innovative way to communicate with public in an informal way. Such method would not substitute the traditional channels of

communication but support them. Also, e-government was and still uses the one-direction of communication between the governments and the public [22]. Public organizations budget cannot afford the ever-evolving technology updates, and the increase in citizens' expectation. This makes governments start searching for innovative ways to deliver their services to citizens. The advent of social media and ubiquities of different ways to connect with different stakeholders helped governments not only to find a channel to disseminate information to public but also to engage public to participate in a political discourse [17].

Mergel [21] discusses the influence that led governments to adopt social media in communicating with public, and found that governments are influenced by four factors: Firstly, they noticed that citizens are using social media. Governments noticed that citizens are using social media to retrieve information and news related to governments, instead of using the official government website. Secondly, the passive observation of agencies that used highly innovative techniques; the intention to adopt social media in government was surrounded with many fears of uncertainty to what extent social media will be accepted. In addition to that governments want to determine the most suitable media for communication. These factors lead governments to observe public and private agencies using social media in communication passively instead of directly communicating with them to understand their experience in this field. Governments try to understand, use and test social media to find the suitable and effective media for their departmental purposes and environment.

Thirdly, the interaction with peers in government agencies, where most governmental systems in the United States of America (USA) are highly centralized, and all headquarters are localized in one location; this enabled the social media director to create a community that enabled the peers to communicate face-to-face or create phone calls weekly in an informal way. Fourthly, formal guidance of lead agencies; the best source of guidance in using social media in public sector comes from major mistakes and the technological change or change in local behavior of specific social media platforms. The other source of guidance came from the president of USA for achieving the triple goals of transparency, participation and collaboration. Also, the existence of social media director in all governmental organization plays a driver factor to success such as adoption.

III. RESEARCH METHODOLOGY

This study proposed a framework for investigating the success of communication between governments and other stakeholders utilizing social media. The framework is founded on a set of proposed factors that lead to communication success and they are: transparency, participation, collaboration, comfort, and the posted topic. This study's major contribution is to sum the factors (based on the literature) and then provides measures and metrics for each factor in the proposed model. Figure 1 depicts our proposition for such environment and the relationships assumed.

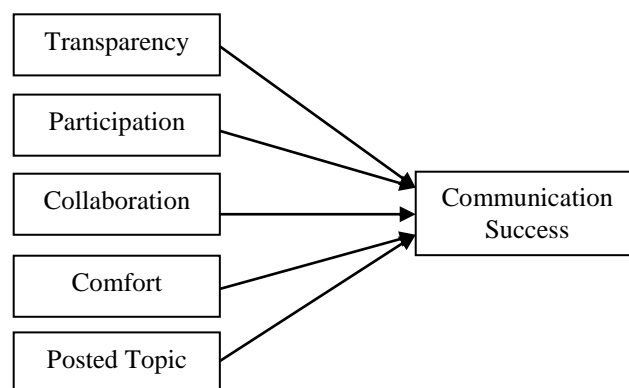


Figure 1. Framework for Government's Communication Success over Facebook

Transparency, governments should make the information as public assets, to enable their stakeholders be intellectual in governments discourse [21]. As we mentioned empirical test on the level of countries supported the relationship between the success of transparency and e-government level of development [1].

Participation, governments should enable stakeholders participate in government topics by opening the dialogue between citizen and governments [27] [4], by enabling their stakeholders to post on their pages or use surveys to collect information from stakeholders [19] [21]. Web 2.0 tools were extensively proposed by Abu-Shanab and Al-Dalou' to enrich and support all levels of e-participation, which leads to better communication and eventually the success of e-government [3].

Collaboration, between governments and other stakeholders, collaboration enable government to benefit from stakeholders experience, knowledge and opinions. Collaboration is one of the highest levels in participation [16] and a major dimension of open government [2] because the government asks other stakeholders to co-create or co-design one of the processes or services and its look like a collective process [11].

Comfort, the use of social media by governments in communicating with its stakeholders for publishing information, provision of services, or getting feedback from them attains the comfort level required from e-services. When governments use the favorite community channel for each stakeholder they will make the communication more comfortable rather than forcing citizens to visit the physical

agency whenever they wanted to get a service or a piece of information [7].

Posted topic, government posted topics may affect the communication between government and its stakeholders. It is noticeable that some of the topics may increase the rate and intensity of communication between the government and stakeholders and some of topics may not attract/encourage stakeholders to communicate.

Communication success, it is so difficult to measure because it depends on the goal of communication via social media. Using available and obvious measures like: the number of likes, comments, shares and reply will provide an indication of polarity of this page and will provide also an indication of stakeholders' engagement with government's activities via social media such as Facebook [11]. As mentioned, the examples of measures are related to Facebook (the focus of this paper), but other measures can be proposed for other types of social media.

a. Measures and Metrics

This section provides measures and metrics for the factors depicted in the previous model (in Figure 1). To measure communication success between a set of governmental Facebook pages we can take a set of samples like the Facebook pages of e-government websites for Arab countries. The focus of measures needs to investigate the following factors: transparency, participation, collaboration and comfort over Facebook pages. Table I to Table V lists the dimensions of each factor in our proposed model (summarized from the literature) and their associated measures.

Table I: Transparency measures and metrics

#	Measure	Metric
1	Number of published posts	Number of published posts ÷ maximum number of posts in sample
2	Number of posts that include calendar events	Number of posts that include calendar events ÷ maximum number of posts that include calendar in sample
3	Number of post that include multi-media:	
	Photos	Number of posts that include photos ÷ maximum number of posts that include photos in sample.
	Videos	Number of posts that include videos ÷ maximum number of posts that include videos in sample.
	Audios	Number of posts that include Audios ÷ maximum number of posts that include Audios in sample.
	Other social media posts	Number of posts that include social media ÷ maximum number of posts that include social media in sample.

Table II: Participation measures and metrics

#	Measure	Metric
1	Government website enables their stakeholder to post over the page	If the government enables its stakeholders to post, a value of 1 will be allocated, else will allocate zero.
2	Number of posts that include survey tool.	Number of posts that include survey tool ÷ maximum number of posts that include survey tool in sample.

Table III: Collaboration measures and metrics

#	Measure	Metric
1	Number of posts that request citizens to engage.	Number of posts that request citizens to engage ÷ maximum number of posts that include a request to engage in sample.
2	Number of posts that request from stakeholders to co-create	Number of posts that request co-creation (design) ÷ maximum number of posts that include co-creation in sample.

Table IV: Comfort measures and metrics

#	Measure	Metric
1	Presences of link to e-government website	if the a link exist, a value of 1 will be allocated, else will take zero
2	Number of posts that include a link to e-government website	1- (Number of posts with a link to e-government website ÷ maximum number of posts that include link to government website in sample).
3	Number of posts that have a link to external website.	1- (Number of posts with a link to external website ÷ maximum number of posts that include link to external website in sample).
4	Number of posts that include application started on Facebook and redirected to e-government website	1-(Number of posts that include posts with application started on Facebook and redirected to e-government web site ÷ maximum number of post that include application redirected to e-government website in sample).
5	Number of posts that include application started and completed in Facebook page	Number of posts that include application started and completed in Facebook page ÷ maximum number of posts that include application started and completed in Facebook website in sample

Table V: Communication success measures and metrics

#	Measures	Metric
1	Like	Total likes for each Facebook page ÷ total followers for each Facebook page
2	Comment	Total comments for each Facebook page ÷ total followers for each Facebook page
3	Share	Total shares for each Facebook page ÷ total followers for each Facebook page
4	Reply	Total Replies for each Facebook page ÷ total

	followers for each Facebook page
--	----------------------------------

b. Measurement

To measure each governmental Facebook page based on the previous tables a thorough inspection of each e-government page on Facebook should be done. As an example, take the transparency measure, first, the total posts of each governmental Facebook page should determine and then divided by the maximum total post in a sample. Then the posts that include calendar event should be estimated and divided by the maximum posts that include calendar events in sample. Also, each multimedia type should be estimated and divided by the maximum post in the sample according to the corresponding type as determined previously.

The next step is to calculate a total measure for transparency construct. Each governmental Facebook page will get an index for transparency by calculating all metrics of transparency and divide by 6 (the number of transparency measures).

To measure the communication success for transparency we take the posts, likes, comments, shares and reply on posts that have the characteristic of transparency that should be counted for each Facebook page and divide by the total followers for each Facebook page.

For example, to calculate the communication success on transparency we take the communication on total posts which is the total (likes, comments, shares, and replies) and divide by the total followers. The following step is to take the communication on total calendar events which is the total (Likes, comments, shares and replies) and divide by the total followers. Also, the communication on total posts that include multimedia which is the total likes, comments, shares and replies) divided by the total followers. To get the communication success over the transparency index it will include the likes, comments, shares and replies) divided by total followers for all transparency measures.

Similarly, when calculating the transparency index and their communication success of the entire sample we can notice that the higher the transparency index, the higher the communication success.

Proposition 1: The higher the transparency index, the higher the communication success

The previous calculations should be conducted for participation and collaboration and get indices for them. So the participation index will be calculated based on all the participation metrics/2 (as we proposed 2 measures). Similar calculations can be done for collaboration index. The following propositions are stated:

Proposition 2: The higher the participation index, the higher the communication success

Proposition 3: The higher the collaboration index, the higher the communication success

On the other hand, the comfort metrics should be calculated in a different way because the comfort of using Facebook should include all the texts and content in the same post without directing the stakeholders into different locations. Based on that, the last metric of comfort is calculated in a different way; because all the content is in the same Facebook page.

Also, about the posted topic (which did not include a table of measures for it); this study did not provide measures and metrics because it can be calculated by classifying the posted topics by governments over Facebook, and find the frequency for each topic and the level of communication for each topic then notice if some topics has higher communication than others. Using this measure may get a good indicator about the most important topics for stakeholders and encourage governments to focus on the most attractive and important topics. The following propositions are stated:

Proposition 4: The higher the level of comfort index, the higher the communication success

III. CONCLUSIONS AND FUTURE WORK

This study explored the literature to better understand the environment of social media and its utilization in e-government communication success. The importance of social media as a communication channel, and the reasons that lead governments to adopt it in their communication are also investigated in the literature section. Also, some models of e-government based on social media are reviewed and the strategies of governments adopted to communicate over the social media.

The major contribution of this study was to propose a framework that included five major predictors of e-government communication success and they are: transparency, participation, collaboration, comfort, and posted topic. The authors also proposed a set of measures and metrics to estimate each factor and conclude to an index for each factor and related to Facebook tools and applications.

These measures and metrics may be used in the future in classifying governmental official Facebook pages based on transparency, participation, collaboration and comfort. Also, an index for each page can be estimated. Finally, the framework proposed here can be used to conduct a comparative empirical study that compare governmental Facebook pages, and also measure an absolute index. Posted topics can be used to uncover what are the topics that attract stakeholders to encourage governments to focus on them.

Researchers are encouraged to apply this method and empirically test our propositions on e-government websites founded on social media. Also, future research can reflect back on our proposed model (if we need other factors to predict communication success), and our proposed measures and metrics (to validate and improve for other researchers.)

REFERENCES

- [1] Abu-Shanab, E. (2013). The Relationship between Transparency and E-government: An Empirical Support. IFIP

e-government conference 2013 (EGOV 2013), September 16-19, 2013, Koblenz, Germany, pp. 84-91.

[2] Abu-Shanab, E. (2015). Open Government Initiatives In Public Sector: A Proposed Framework For Future Research. *Saba Journal of Information Technology and Networking*, Vol. 3(1), pp. 4-14.

[3] Abu-Shanab E. & Al-Dalou', R. (2012). E-participation Initiatives: A Framework for Technical Tools, *The 2012 International Arab Conference of e-Technology (IACe-T'2012)*, Zarqa, Jordan, April 25-27, 2012, pp. 57-64.

[4] Al-Dalou', R. & Abu-Shanab, E. (2013). E-Participation Levels and Technologies. The 6th International Conference on Information Technology (ICIT 2013), 8-10 May, 2013, Amman, Jordan, pp.1-8.

[5] Al-Saggaf, Y. & Simmons, P. (2014). Social media in Saudi Arabia: Exploring its use during two natural disasters. *Technological Forecasting & Social Change*. pp.1-13.

[6] Arab Social Media Report. (2014). Citizen Engagement and Public Services in the Arab World: The Potential of Social Media, a report published by Mohammed Bin Rashid School Of Government.

[7] Camacho, R., & Kumar, M. (2012). Social Media on e-Government. Accessed in 2014 from: http://www.academia.edu/1958732/Social_Media_on_e-government), visited on 29/11/2014.

[8] Chun, S., & Luna Reyes, L. F. (2012). Social media in government. *Government Information Quarterly*, vol. 29(4), pp. 441-445.

[9] Diamond, L. (2010). Why are there no Arab democracies? *Journal of Democracy*, vol. 21(1), pp.93-112.

[10] Elefant, C. (2011). The "Power" of Social Media: Legal Issues & Best Practices For Utilities Engaging Social Media. *Energy Law Journal*, Vol.32 (1).

[11] Hofmann, S., Beverungen, D., Räckers, M., & Becker, J. (2013). What makes local governments' online communications successful? Insights from a multi-method analysis of Facebook. *Government Information Quarterly*, vol. 30(4), pp.387-396.

[12] Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons*, vol. 53(1), pp.59-68.

[13] Khasawneh, R. & Abu-Shanab, E. (2013). E-Government and Social Media Sites: The Role and Impact. *World Journal*

of Computer Application and Technology, Vol. 1(1), July 2013, pp. 10-17.

[14] Khasawneh, S., Jalghoum, Y., Harfoushi, O., & Obiedat, R. (2011). E-Government Program in Jordan: From Inception to Future Plans. *International Journal of Computer Science Issues(IJCSI)*, vol. 8(4), pp.568-582.

[15] Klischewski, R. (2014). When virtual reality meets realpolitik: Social media shaping the Arab government-citizen relationship. *Government Information Quarterly*, vol. 31(2), pp.358-364.

[16] Lee, G., & Kwak, Y. H. (2012). An Open Government Maturity Model for social media-based public engagement. *Government Information Quarterly*, vol. 29(4), pp.492-503.

[17] Linders, D. (2012). From e-government to we-government: Defining a typology for citizen coproduction in the age of social media. *Government Information Quarterly*, vol. 29(4), pp.446-454.

[18] Magro, M. J. (2012). A review of social media use in e-government. *Administrative Sciences*, vol. 2(2), pp.148-161.

[19] McDermott, P. (2010). Building open government. *Government Information Quarterly*, vol. 27(4), pp.401-413.

[20] Mergel, I. (2013 a). Social media adoption and resulting tactics in the US federal government. *Government Information Quarterly*, vol. 30(2), pp.123-130.

[21] Mergel, I. (2013 b). A framework for interpreting social media interactions in the public sector. *Government Information Quarterly*, vol.30(4), pp.327-334.

[22] Mossberger, K., Wu, Y., & Crawford, J. (2013). Connecting citizens and local governments? Social media and interactivity in major US cities. *Government Information Quarterly*, vol.30(4), pp.351-358.

[23] Picazo-Vela, S., Gutiérrez-Martínez, I., & Luna-Reyes, L. F. (2012). Understanding risks, benefits, and strategic alternatives of social media applications in the public sector. *Government information quarterly*, vol.29(4), pp.504-511.

[24] Salih, K. E. O. (2013). The roots and causes of the 2011 Arab uprisings. *Arab Studies Quarterly*, vol.35(2), pp.184-206.

[25] Snead, J. (2013). Social media use in the U.S. Executive branch. *Government Information Quarterly*, vol.30 (5), pp. 56-63.

[26] Storck, (2011). The Role of Social Media in Political Mobilisation: a Case Study of the January 2011 Egyptian Uprising. Accessed at (http://www.culturaldiplomacy.org/academy/content/pdf/participant-papers/2012-02-bifef/The_Role_of_Social_Media_in_Political_Mobilisation_-_Madeline_Storck.pdf), visited at 29/11/2014.

[27] Zavattaro, S. M., & Sementelli, A. J. (2014). A critical examination of social media adoption in government: Introducing omnipresence. *Government Information Quarterly*, vol.31(2), PP.257-264.

A Review on Internet Banking Security and Privacy Issues in Oman

Elbek Musaev

Department of Management Information Systems
Dhofar University
Salalah, Sultanate of Oman

Muhammed Yousoof

Department of Management Information Systems
Dhofar University
Salalah, Sultanate of Oman

Abstract— Internet banking (IB) is not a new phenomenon anymore as more and more financial institutions worldwide jump onto this wagon as it creates win-win situation for all parties. There is no need to go to bank office to pay bills, check account balance and make funds transfer. Today banks with significant IB experience provide even more complicated online financial tools and services. Nonetheless, due to the fact that platform of IB is World Wide Web, security and privacy issues are of high concern. So banks in Oman, lacking technically advanced experience of other countries should provide more safe and secure IB services, as security issues in this vulnerable area do exist. This work studies security and safety problems and suggests theoretical and practical recommendations.

Keywords— *Internet banking; security; privacy; mobile banking*

I. INTRODUCTION

Due to rapid development of interconnected online IT infrastructure financial institutions around the world urged to keep up with this development as many see the future of commerce and affairs done online. People can do banking operations sitting home, at work, or lying on their beds midnight as this can be done through computers or mobile devices. Internet Banking (IB) was defined as distantly performing financial transactions over internet with the help of bank's website [1]. Since banks provide internet-based services, they should have secure and reliable methods of authenticating their customers [2]. Therefore, banks have to better understand their customers, current adoption of IB and respond quickly to market developments by identifying reasons that impact customer perception of security and usability issues in IB [3].

We believe that many banks in Oman have security issues as there was a biggest ATM fraud heist in history of USD 45mln by hackers worldwide. The cash withdrawals were made through ATMs in 24 countries including the US, Germany, Japan, Russia, Romania, Egypt, Colombia, Britain, Sri Lanka and Canada. Hackers accessed Bank Muscat and Rakbank databases, removed withdrawal limits on prepaid debit cards and created access codes. Others loaded that data onto any expired plastic card with a magnetic stripe and distributed among themselves, thus stealing loads of money [4]. This case subsequently might have led to low use of IB in Oman, which can be seen in statistics discussed next.

As Oman is relatively new country that started providing essential provisions to keep up with the modern developments and needs to study from other technically advanced counterparts. According to Information Technology Authority – the institution responsible for implementing of Digital Oman Strategy – 67% of total population has used Internet in 2013, and only 8% of them used IB. This confirmed by additional statistics – a large majority, 85% of internet users have never bought or ordered anything online and only 8.3% conducted E-commerce activities within 3 months. In addition, when using E-government services, only 14% used online services or submitted online forms, the rest either downloaded online forms or just obtained information. Main reason for not using E-government services was no necessity (45%) or concerns about protection and security of personal data 44%) [21].

This leads to suggestion that one of major reasons to low usage of E-commerce and IB in Oman is fear of leakage of private data and security of services provided. Therefore this work in progress will be done in two stages: 1) survey bank customers and see if security issues lead to poor adoption of Internet and mobile banking and 2) find out IB security issues of three major banks in Oman through benchmarks of IB in South Korea and propose solutions. As people start slowly migrating from desktop computers to mobile tablets and smart phone devices, this study will examine both IB access from desktop and Mobile banking (MB) security issues, as well as mobile applications of studied banks will be thoroughly analyzed.

II. INTERNET BANKING

There was a growing trend in the amount and attention given to IB adoption research that have increased over time and will supposedly remain as key area of studies in coming years [5] [18]. IB services give customers possibility most of traditional services without the need of going to bank offices saving time and money of both sides. For the past decade, IB usage has grown substantially and banks worldwide began to give this occurrence more attention and support[6]. Interestingly, another study argues that the “growth rate of those who adopt IB has not risen strongly as expected” [3] as there is little knowledge of true determinants of online banking adoption [7]. Technology acceptance model is used in this study to indentify if security and risk are having a good weight in adoption of IB and therefore can be accepted as major determinants.

In addition, as we see in everyday life that many people move from desktop and even laptop PCs to mobile equivalents like tablet PCs and smart phones. Thus, mobile banking being a subset of IB is likely to be used overwhelmingly and replace IB in the future due to convenience and ease of use of mobile devises [3] [8]. For this reason, large and commercial banks that always diversify their activities in constant search of additional benefits tend to quickly adopt mobile banking, offer more mobile financial services, security features and support more devices [9]. This research will try to identify any available online security issues and give practical recommendations to managers of bank institutions.

III. METHODOLOGY

First part of research was conducted by surveying bank customers and find out if security is a major factor that causes slow usage and adoption of IB. Technology acceptance model (TAM) was used to identify whether insecurity in form of trust and risk influence the usage of IB, because it is one of the most influential theory models that predicts the adoption of any certain technology in Information Systems. TAM argues that perceived usefulness and perceived ease of use define an intention to use a system [5]. This work applies TAM with additional factors that describe security.

A. IB Security Issues

Perceived lack of security is “a perceived potential loss due to fraud or a hacker compromising the security of IB” [10]. So, IB threats can be accomplished through network attacks or through illegal access to the customer account by means of fabricated or faulty authentication [3]. For this reason, security and privacy threats are constantly increasing both in quantity and quality [11]. Thus, “online banking security is a primary concern”, as banks supposedly provide all measures necessary to make customers believe that information is transmitted safely and securely [12]. Awareness of security has direct impact on trust and usage of IB and indirectly affects perceived ease of use [7], in other words security issues have significant effects on IB use [3] having negative causal relationship [10] [13]. To sum up, banks and financial institutions should build and enhance confidence and trust of their internet services and data transfer, as well as provide privacy and security protection, system reliability and financial quality information [14].

Thus we hypothesize that:

H1a: Trust has positive effect on the intention to use Internet banking.

H1b: Trust has positive effect on usefulness of Internet Banking

On the contrary, Lee et al. suggest that in South Korea few cases of customer fraud were reported due to good IB security infrastructure, but “serious potential problem now and in the future is leakage of private information” [15].

Nonetheless, other scholarly research supports ideas opposing security problems, e.g. security is not perceived as an obstacle or a major concern in mobile banking transactions [16]. It can be understood that people are ready to settle with less secure environment in favor of ease of use and convenience. Customers’ IB use satisfaction drops if they have to memorize multiple pieces of credentials and use One-time password tokens provided by banks [7]. Therefore majority of people use simple, easy-to-remember, “convenient” passwords that can simply be guessed by people. SplashData provides such list of easily guessable, most frequently used passwords [17]. Thus people can trade-off encrypted, more secure Android phone to non-encrypted, as fully encrypted Android device, due to security, privacy concerns allows locking phone only with at least 6 character password, disabling pin, face, pattern and fingerprint unlocks. Thus it becomes very inconvenient to type passphrase every time to unlock a phone. Hence you can see many people use simple 4 digit PIN, pattern unlock, face unlock or by fingerprint. Latter two are biometric types of unlock and are a good research directions for future. Therefore usefulness and ease of use are critical factors in TAM:

H2a: Usefulness positively influences the intention to use Internet banking

H2b: Usefulness positively influences the ease of use of Internet Banking

H3: Ease of use has positive effect on intention to use Internet banking

Perceived risk is another major concern when customers connect remotely to bank servers and do various transactions. Thus it was found that perceived risk has considerable effect on intention to use [22] as risks are considered as important factors for frequent or non-frequent users [23].

H4: Perceived risk has negative effect on intention to use Internet Banking

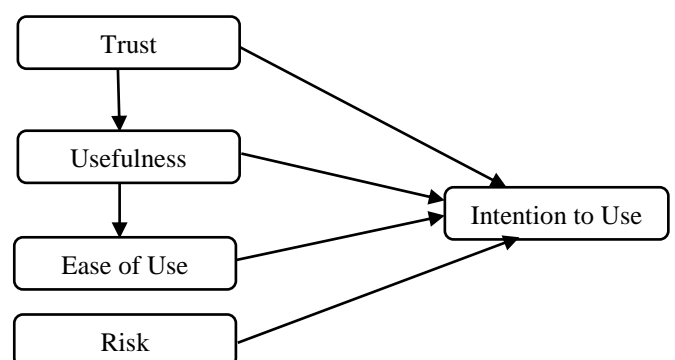


Fig.1 Hypothesized model

B. Technical Analysis and Interview

Second part of the research will verify the online security of three big banks: Bank Muscat, Bank Dhofar and HSBC as a representative of foreign experience. The security of online banking application will be concentrated on three levels [12], security threats – measures approach and will further be extended from experience of South Korean banks. Security threats – measures approach includes several types of threats and measures – internal, external, human, non-human, accidental and intentional categories [19].

- *Security of customer information sent from PC or mobile device to web server.*

This includes the availability of secured website information of bank, namely IB section should be always be 128-bit and above secure socket layer (SSL) encrypted in order to remove man-in-the-middle external threats. However, such method is not proven to be completely secure as man-in-the-middle attack by hackers may lead to personal information and credentials leak [2] or people can use insecure ways of storing their login credentials on a piece of paper after forgetting passwords, login IDs, secret question – answers [19]. For this reason, if hackers use key logger software to gather information, website should include option of virtual input of data with the help of virtual keyboard. Study of two-factor authentication (2FA) of UK banks has also found that some websites do not hide passwords or pass-phrases [6]. In addition, from the personal experience of IB in South Korea, most of IB of Korean banks initially use additional software like anti-virus, anti-key logger, anti-screen capture, anti-malware installed obligatory before entering into IB section. After first installation, the next time you log in, these programs automatically run in the background to secure connection and transfer of any data from user PC to bank web server.

This method of installing additional software seem to have proven itself with time, as study of security of IB and financial private information in South Korea indicates that there were few cases of customer hacking fraud was reported [15].

- *Security of web server environment and customer information database.*

Real world example has shown that even banks and financial institutions pay great attention and invest heavily in security issues, all information systems have weaknesses that create opportunities for possible threats to the information housed in these systems [19]. As a result, Lee et al. suggest using central government regulated encrypted repository of all customer private information with the help of e-pin [15]. This means that after first registration in any bank, customer information is sent to the central repository and after authenticating user a unique password-protected e-pin number is issued. Next registration in bank will not require user to enter personal data again, but to enter the unique e-pin and the data will be retrieved by bank automatically and verified. This method will effectively eliminate perceived information system weaknesses and reduce possibility of external attacks.

- *Security measures to prevent unauthorized access to IB section.*

Two-factor authentication (2FA) has proven to be secure method customer verification requiring them to produce

additional authentication [6] [11] together with their unique login ID and password like one time passcode (OTP) issued by OTP token or received by SMS to mobile device, phone call, card reader or a card with random numbers. Nonetheless, increased security could have negative effect on the IB system use [19] such are too many specific information, predefined security questions and answers that needed to be remembered by customers to verify their entry into IB section lead to decreased perceived usability [5] [6] [18] [19]. Moreover, 2FA schemes have conceptual vulnerabilities and not completely secure because OTPs can be intercepted [11].

Among other security measures are: session timeouts or auto-terminal/account logoff, automatic lockouts after a number of unsuccessful login tries, use of strong passwords [19].

Thus, security issues discussed above can be summarized as follows:

1. *Security of customer information sent from PC to web server.*
 - a) Are online sessions secured at all times, i.e. are all web pages interacting with online banking 128(256) bit SSL encrypted (registration, login, fund transfer pages).
 - b) Type passkey, passphrase, user ID by virtual keyboard (not physical keyboard).
 - c) Entered characters hidden (e.g. asterisk *)
 - d) Requirement of additional security software installed before registering or using online banking section (anti-virus, anti-spyware, anti-key logger, anti-screen capture).
2. *Security of web server environment and customer information database.*
 - a) Is customer information saved in bank web server?
 - b) Is customer information database securely protected?
3. *Security measures to prevent unauthorized access to online banking section.*
 - a) Does bank use two-factor authentication (2FA)?
 - One time password token carried by customer
 - Card reader
 - A card with random numbers issued by bank
 - Mobile phone SMS authentication
 - Phone call
 - b) Does bank use 2FA after login and for transactions?
 - c) Does bank use 2FA only for transactions?
 - d) Are there session timeouts after some time?
 - e) Do automatic lockouts after 3 unsuccessful tries exist?
 - f) Is creation of strong passwords suggested and assisted on web page?

IV. RESULTS OF SURVEY

The survey was conducted in Salalah, Oman from the sample of 200 respondents, returned results were 121 and after

removing 14 unusable we had 107 filled survey papers at hand to analyze the model. 55% of respondents were females and 45% – males. 75% are at the age of between 20-29, 60% of respondents are Muscat bank users, 24% - Dhofar and few numbers of HSBC and National Bank of Oman. Largest share almost 70% were students.

The regression and correlation analysis was used to check the model and relationship of variables.

TABLE 1. CORRELATION RESULTS OF VARIABLES

	Intention to Use IB
Trust in IB	.914**
IB Usefulness	.747**
Ease of Use	.776**
Risk	.385**

** - Correlation is significant at the 0.01 level (2-tailed)

In addition, the correlations between trust, usefulness and ease of use are significant. Trust and usefulness - .741, usefulness and ease of use - .736. Other two correlations we didn't consider also have shown to be significant – trust and ease of use (.749) as well as risk and usefulness (.603). All correlations are significant at 10% level. Thus we can understand that it is paramount for that bank to gain trust of customers in order to introduce and encourage to use IB by providing safe and secure online services.

By testing the model by the analysis of variance we can understand that independent variables have good weight in explaining the model and intention to use IB.

TABLE 2. STATISTICAL INDICATORS OF THE MODEL

R	R ²	St.Err	Mean.	Chronbach	Sig.
.932	.869	.876	85.4	.902	.000

TABLE 3. HYPOTHESES TEST RESULTS

	Mean	St.dev.	Sig.	Hypothesis
H1a	4.03	2.06	.000	Accept
H1b	4.52	1.97	.004	Accept
H2a	4.58	1.91	.000	Accept
H2b	4.56	1.64	.000	Accept
H3	4.06	2.02	.000	Accept
H4	4.07	1.75	.003	Accept

Sig. level <0.05

Results have shown that all of the hypotheses are accepted at various significant levels.

V. DISCUSSION AND IMPLICATIONS

The analysis has shown the expected results of trust, usefulness, ease of use and risk having various degrees of influence on the intention to use IB.

Any research has limitations, as current work was based on student survey and in Salalah. The study of respondents in Muscat will certainly show different results as they live closer to Dubai, world trade hub and may use IB more often than those who live in Salalah. This is due to the fact that most people prefer traveling to Dubai directly and do their shopping there. Creating e-commerce presence between Dubai and Omani shoppers would greatly help reduce traveling costs and time. Thus by using more e-commerce people would use IB more and other forms like MB.

As the chosen variables not completely determine the intention to use IB, more research should be conducted in order to find all determinants of the model implemented in this work.

As to implications to practitioners, banks should develop trust of customer perceptions in use of new technologies and review security features of website and mobile applications. They are advised to pay attention to make the use of services easy and useful by explaining and showing that customers can save time by doing non-cash services. Risk from losing money by possible attack of hackers was supported weakly which means that customers care mostly about trust and ease of use, together with usefulness.

Second part of the work will be conducted later in order to get whole picture of bank security issues in Oman, however we can derive following by looking into some features:

Security of customer information sent from PC to web server.

Study showed that one of reviewed bank website login page is secure and encrypted and has virtual keyboard input, but when after navigating to bank registration page, web page vulnerability was detected by internet browser indicating that data is not completely secure and can be intercepted by external human factor that could intentionally steal required information. Thus, private data is not secured and after stealing that information, hackers have better chances to access IB web section with little or no efforts [19].

If one compares this case with website of South Korean banks, it can be seen that when entering to IB section, website forces to install additional software like anti-virus, anti-malware, anti-key logger and anti-screen capture. This ensures that private data and credentials are secure when transmitting data from user PC to web server. Therefore we can conclude that there were not many online cases of money fraud in South Korea [15].

Detailed study will be conducted later with the interviews of bank representatives and analysis of the websites and mobile applications of related banks.

Security of web server environment and customer information database.

Security of bank web server environment is not completely secure and even high tech South Korea banks were often hacked by international cyber crime groups, however there was no case in Korea like with Oman. There are two possible solutions for Omani banks to possibly prevent breach of web

servers. Firstly, all banks should carry out independent audit of their information systems by network security analysts or hackers [19]. Secondly, as it was suggested earlier, use of centralized repository of population of Oman in one place with the help of e-pins [15] for the reason that inter-institution networks are more protected and secured and it is easier to protect one system effectively, than trying to secure many places at once.

Security measures to prevent unauthorized access to online banking section.

All of the banks reviewed use 2FA security features with HSBC providing physical secure keys. Other banks use SMS based 2FA. However, these banks use 2FA for money transaction only, and not for login sessions. Although banks in South Korea also use 2FA for transactions only their IB environment is more secure than those of Oman. For other security features like session timeouts and lockouts all banks met the requirement. However, they do not give customers feedback to check whether passwords are strong or not, but only mention about use of strong passwords.

Many studies suggest use of biometrics [6] [2] as a solution to complicated process of memorization of specific data and 2FA replacement. Among the suggestion is voice recognition [9], hand eye scan and fingerprint [2]. Saleh suggests using RFID to authenticate customers in order to improve IB security and improve trust [12]. Thus scholarship research could give more attention to research novel ways of authenticating IB customers.

REFERENCES

- [1] G. Shao, "The diffusion of online banking: research trends from 1998 to 2006", *Journal of Internet Banking and Commerce*, vol. 12, No. 2, August 2007, pp. 1-13.
- [2] F. Amtul, "E-Banking security issues – is there a solution in biometrics?", *Journal of Internet Banking and Commerce*, vol. 16, No. 2, August 2011, p.1.
- [3] H.S. Yoon and L. Occena, "Impacts of customers' perceptions on internet banking use with a smart phone", *Journal of Computer Information Systems*, vol. 54, No. 3, Spring 2014, pp. 1-9.
- [4] B. Thomas, "9 arrested for \$45mn bank muscat, rakbank prepaid card fraud", <http://www.muscatdaily.com/Archive/Oman/9-arrested-for-45mn-bank-muscat-Rakbank-prepaid-card-fraud-290x>, (Accessed in February 2015), May 2013.
- [5] P. Hanafizadeh, B.W. Keating and H.R. Khedmatgozar, "A systematic review of Internet banking adoption", *Telematics and Informatics*, Vol. 31, No. 3, April 2014, pp. 492-510.
- [6] K. Krol, E. De Cristofaro and A. Sasse, "They brought in the horrible key ring thing!" Analysing the usability of two-factor authentication in UK online banking", Cornell University Library, [arXiv:1501.04434](https://arxiv.org/abs/1501.04434), unpublished
- [7] M.S. Alnsour and K. Al-Hyari, "Internet banking and Jordanian corporate customers: Issues of security and trust", *Journal of Internet Banking and Commerce*, vol. 16, No. 1, April 2011, p.1.
- [8] R. Weber and A. Darbellay, "Legal issues in mobile banking", *Journal of Banking Regulation*, vol. 11, No. 2, 2010, pp. 129-145.
- [9] H. Lee, Y. Zhang and K.L. Chen, "An investigation of features and security in mobile banking strategy", *Journal of International Technology and Information Management*, vol. 22, No. 4, October 2013, pp. 23-46.
- [10] M. Lee, "Factors influencing the adoption of Internet banking: An integration of TAM and TPB with perceived risk and perceived benefit", *Electronic Commerce Research and Applications*, vol. 8, No. 3, May-June 2009, pp. 130-141.
- [11] A. Dmitrienko, C. Liebchen, C. Rossow and A.-R. Sadeghi, "Security analysis of mobile two-factor authentication schemes", *Intel[®] Technology Journal*, vol. 18, No. 24, 2014, pp. 138-161.
- [12] Z. Saleh, "Improving security of online banking using RFID", *Academy of Banking Studies Journal*, vol. 10, No. 2, July-December 2011, pp. 1-8.
- [13] C. Kim, M. Mirusmonov and I. Lee, "An empirical examination of factors influencing the intention to use mobile payment", *Computers in Human Behavior*, vol. 26, No.3, May 2010, pp. 310-322.
- [14] J-P.L. Mangin et al., "The moderating role of risk, security and trust applied to the TAM model in the offer of banking financial services in Canada", *Journal of Internet Banking and Commerce*, vol. 19, No. 2, August 2014, pp. 1-21.
- [15] J.H. Lee, W.G. Lim and J.I. Lim, "A study of the security of Internet banking and financial private information in South Korea", *Mathematical and Computer Modeling*, vol. 58, No. 1-2, July 2013, pp. 117-131.
- [16] T. Laukkanen, "Internet vs mobile banking: comparing customer value perceptions", *Business Process Management Journal*, vol. 13, No. 6, 2007, pp. 788-797.
- [17] "Password" unseated by "123456" on SplashData's annual "Worst Passwords" list" retrieved from <http://splashdata.com/press/worstpasswords2013.htm>, (accessed in February 2015).
- [18] H.M. Sabi, "Research trends in the diffusion of Internet banking in developing countries", *Journal of Internet Banking and Commerce*, vol. 19, No. 2, August 2014, pp. 1-31.
- [19] A. French, "A case study on E-Banking security – When security becomes too sophisticated for the user to access their information", *Journal of Internet Banking and Commerce*, vol. 17, No. 2, August 2012, pp. 1-14.
- [20] F. Sidi et al., "Measuring computer security awareness on Internet banking and shopping for Internet users", *Journal of Theoretical and Applied Information Technology*, vol. 53, No. 2, July 2013, pp. 210-216.
- [21] "Survey on access to, and use of ICT by households and individuals in Oman 2013", retrieved from http://www.ita.gov.om/ITAPortal/MediaCenter/Document_detail.aspx?NID=97, (accessed in January 2015)
- [22] A. Kesharwani, S.B. Singh, "The impact of trust and perceived risk on internet banking adoption in India: An extension of technology acceptance model", *International Journal of Bank Marketing*, Vol. 30, No. 4, 2012, pp. 303-322.
- [23] C. Chen, "Perceived risk, usage frequency of mobile banking services", *Managing Service Quality*, Vol. 23, No. 5, 2013, pp. 410-436.

Information and Knowledge Engineering



A survey on Applying Ontological Engineering Approach for Hepatobiliary System Diseases

Galal AL Marzoqi, Marco Alfonse, Ibrahim F. Moawad, Mohamed Roushdy

Faculty of Computer and Information Science, Ain Shams University, Abbasia, Cairo, Egypt
galalalmarzoqi@gmail.com, marco@fcis.asu.edu.eg, ibrahim_moawad@cis.asu.edu.eg, mroushdy@cis.asu.edu.eg

Abstract— Medical Ontologies play a central role in integrating heterogeneous databases of various model organisms. Hepatobiliary system is very important to human vital processes. It has an ability to regulate the other systems. Furthermore, it may be affected by many pathologic conditions, which affect other organs negatively. This paper investigates the current studies on Ontological engineering approach and Ontology techniques for Hepatobiliary System Diseases. We present conceptual view for the Hepatobiliary system and its infected diseases. Besides, we propose a new classification schema for the research efforts investigated so far. We classified the research efforts investigated so far based on the Hepatobiliary system organs: Liver, Gallbladder, Bile duct and Pancreas. Besides, we discuss the current research gaps found in this research area.

Keywords— *Hepatobiliary System Diseases; Ontology Engineering; Protégé; Medical Systems;*

I. INTRODUCTION

Ontology is a kind of controlled vocabulary of well defined terms with specified relationships between those terms, capable of interpretation by both humans and computers [1]. Furthermore, it is a specific rich description of Terminology, Rules, Concepts, and Relations among the concepts. There are several Ontology languages such as Extensible Markup Language (XML) [2], Resource Description Framework Schema (RDF(S)) [3], Darpa Agent Markup Language Ontology Interface Language (DAML+OIL) [4], and Web Ontology Language (OWL) [5]. The Ontologies can be exploited in many applications in fields, where semantics-based communication among people and systems are crucial. [6]. There are different techniques related to the Ontologies: Ontology alignment, Ontology mapping/matching, Ontology translation, Ontology merging/integrating Ontology refinement and Ontology unification [7].

Ontology tools can be applied for all stages of the Ontology lifecycle (creation, population, validation, deployment, maintenance, and evolution), and hence there are many tools for Ontologies management in different formats (Protégé, OilEd, Apollo, RDFedit, OntoLingua, OntoEdit, WebODE, KAON, ICOM, DOE, and WebOnto) [8]. Ontology can be used to support various knowledge management issues including knowledge retrieval, storing, and sharing [9]. Protégé is an open source software that provides user community with a suite of tools to construct domain models and knowledge-based applications with Ontologies Protégé implements a rich set of knowledge-modeling structures and actions that support the creation, visualization, and manipulation of Ontologies in various representation formats. Protégé can be customized to provide domain for creating knowledge models and entering data [10]. Medical Ontologies

play a central role in integrating heterogeneous databases of various model organisms and

stored in heterogeneous databases. Also, it defines a precise and shared vocabulary for the semantic markup of resources and their description by metadata [11]. Conceptually, it is interested in solving important issues such as the reusing and sharing of medical data. The unambiguous communication of complex and detailed medical concepts is now a crucial feature of medical information systems [12]. There are Medical Ontologies developed to facilitate this purpose such as (Open Biomedical Ontologies OBO [13], National Center for Biomedical Ontology NCBO's BioPortal [14] and Unified Medical Language System UMLS [15]). Medical Ontology is now widely acknowledged that Ontologies can make a significant contribution to the design and implementation of information systems in the medical field.

On the other hand, human body systems consist of specific cells, tissues, and organs that work together to perform specific functions. Conceptually, these systems are interconnected and dependent so they can't work separately such as (Nervous System, Respiratory System, Immune System, Digestive System and Hepatobiliary System) [16,17]. The Hepatobiliary system includes four organs (Liver, Gallbladder, Bile duct and Pancreas). Where, Hepatobiliary is the one of the important systems in human body. It is responsible for lots of processes inside the body. These processes are important to keep body regulated and healthy. Conceptually, it plays an important role in many body functions like protein production. It is also responsible for detoxification, metabolism, synthesis, and storage of various substances however, this vital system may be affected by hazardous conditions whether they are internal or external [18,

19]. Interestingly, these diseases were classified based on different dimensions such as (Cause, Treatment, Symptoms, etc). This paper presents a new classification schema for the research efforts investigated so far. We classified the research efforts investigated so far based on the Hepatobiliary system organs: Liver, Gallbladder, Bile duct and Pancreas.

The paper is organized as follows. Section 2 presents the conceptual view of Ontology based Hepatobiliary Systems. Section 3 displays Liver based Systems, section 4 presents Gallbladder based System, section 5 presents Pancreas based Systems, and section 6 presents Comparative among Ontology based systems for Hepatobiliary system diseases. Finally, section 7 presents conclusion and future work.

II. CONCEPTUAL VIEW ONTOLOGY BASED HEPATOBILIARY SYSTEMS

In general human body consists of many systems such as (Nervous System, Respiratory System, Immune System, Digestive System and Hepatobiliary System). The Hepatobiliary system includes four organs (Liver, Gallbladder, Bile duct and Pancreas). To understand the research paper so far Hepatobiliary system, we design conceptual view for human body and its related systems. Figure 1 shows a new classification schema in the “Human Body” is the main class. Also we have the “system” class. The “system” class into six subclasses which are “Nervous”, “Respiratory”, “Immune”, “Digestive”, “Hepatobiliary” and “Others”. The “Hepatobiliary” class has “Organ” class. The “Organ” class includes five subclasses: “Liver”, “Gallbladder”, “Bile duct”, “Pancreas” and “Disease”. The “Disease” class includes eight instances: “Viral Hepatitis”, “Liver Cancer”, “Liver Immune”, “Chronic”, “Diabetes”, “Anti-diabetic drugs”, “Hepato-Pancreato-biliary” and “Lithiasic Cholecystitis” which are effect on organs: Liver, Gallbladder, Bile duct and Pancreas in Hepatobiliary system.

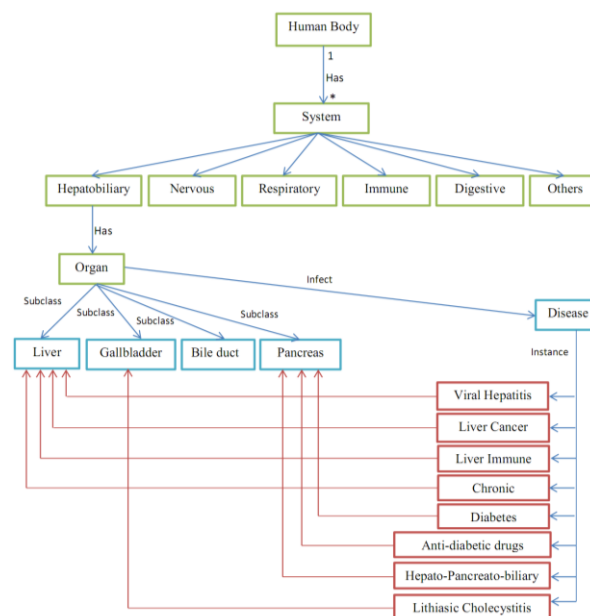


Fig.1. Conceptual view of Hepatobiliary System

There are many domains specific Ontological engineering approach and Ontology techniques has been built on Hepatobiliary system diseases. Figure 2 shows the research efforts investigated so far based on the Hepatobiliary system organs: Liver, Gallbladder, Bile duct and Pancreas.

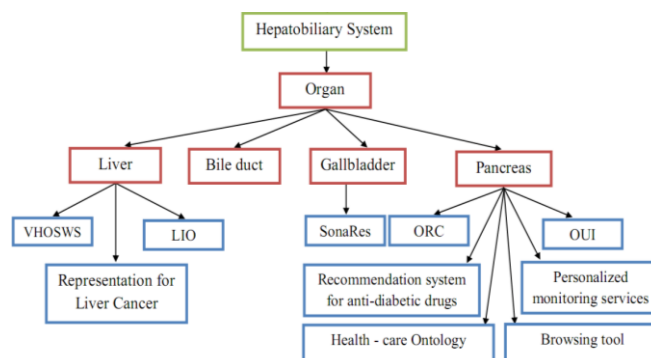


Fig.2. Ontology Based Systems for Hepatobiliary System Diseases

III. LIVER BASED SYSTEMS

In [20], the authors presented a web service based approach to share the Viral Hepatitis Ontology among physicians, students of medicine, and intelligent systems. In addition, the proposed approach enables physicians, and students of medicine to differentially diagnose the Viral Hepatitis diseases. To show how the approach is very beneficial for physicians and students of medicine, the authors developed a system prototype (VHOSWS) to present different usage case studies. In figure 3 the VHOSWS tool for Viral Hepatitis differential diagnosis is shown.

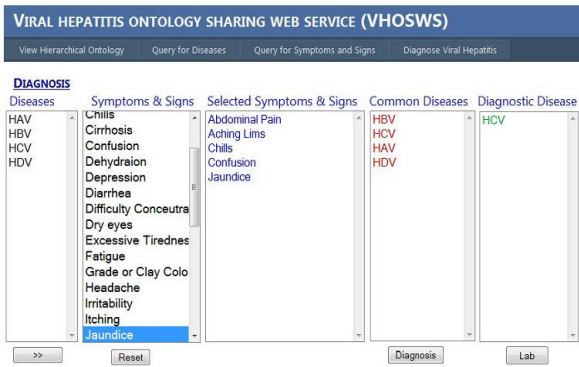


Fig.3. Viral Hepatitis Differential Diagnosis [20]

In [21], the authors developed Ontology based knowledge representation for Liver Cancer that was built using the Protégé-OWL editing environment. It has a great user interface that eases the process of building or editing Ontologies. This Ontology is encoded in OWL-DL format which is the most recent development in standard Ontology languages, endorsed by the World Wide Web Consortium (W3C) to promote the Semantic Web vision.

This Ontology can be used by experts or medical researchers who want the liver cancer knowledge to be represented in a semantic way that allows reasoning capabilities. Figure 4 shows the liver cancer class hierarchy.

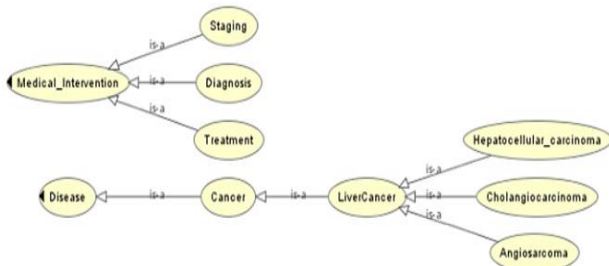


Fig.4. Liver Cancer class hierarchy [adapted from [21]]

In [22], the authors developed Liver Immunology Ontology (LIO) within the Open Biomedical Ontologies (OBO) Foundry framework, importing and linking relevant portions of orthogonal reference Ontologies. LIO is a novel tool for comprehensive analysis of liver immunology data sets, providing a valuable resource for the liver disease research community.

IV. GALLBLADDER BASED SYSTEM

In [23], the authors presented description of the process of Ontology construction for gallbladder ultrasound images. This Ontology is inspired and based on the knowledge base created and being used for SonaRes the decision support system for ultrasound diagnostics. This system has accumulated the experience of the skillful experts-sonographers in the domain of hepato-pancreato-biliary zone examination. This experience and knowledge is well structured and formalized in this system for gallbladder and pancreas.

On the other hand, there is a powerful and attractive, from the point of view of knowledge portability, tool- Ontology, which in computer science is considered as an attempt of comprehensive and detailed formalization of some knowledge domain with the help of conceptual scheme. In figure 5 represents of knowledge on gallbladder pathology.



Fig.5. Knowledge on gallbladder pathology”Chronic lithiasic cholecystitis” [23]

V. PANCREAS BASED SYSTEMS

In [24], the authors presented a method of context-driven annotation in images of the DICOM standard [25] and its application for ultrasound images. Here is described them attempt to create a mapping between the Classification of Diseases ICD-10 [26], and the Ontology of Ultrasound Images (OUI) of hepato-pancreato-biliary zone organs [27].

In [28], the authors developed a disease Ontology based on River Flow Model and a browsing tool for causal chains defined in it. Because the Ontology is based on Ontological consideration of causal chains, it could capture characteristics of diseases appropriately. The definition of disease as causal could be also very friendly to clinicians since it is similar to their understanding of disease in practice. Moreover, it could include richer information about causal relationships in disease than other disease Ontologies or medical terminologies such as SNOMED-CT. Figure 6 presents the types of diabetes constituted of casual chain.

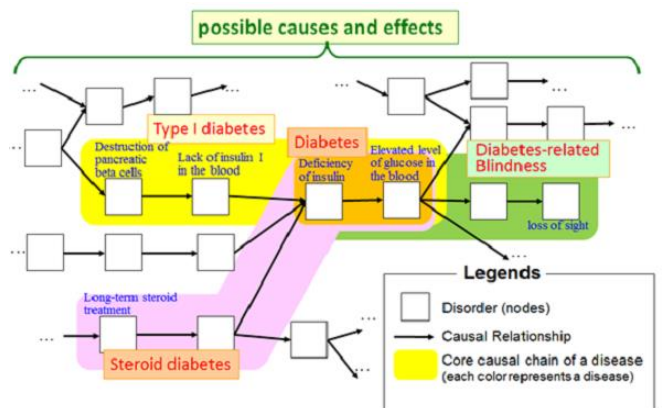


Fig.6. Types of Diabetes constituted of casual chain [28]

In [29], the authors presented an Ontology reasoning component (ORC) that builds upon existing Ontology modeling tools and techniques to support the integration and interpretation of multimodal medical information. He had illustrated how to embed ORC as a reasoning capability in reactive infrastructure agents supporting intelligent agents operating in COMMODITY₁₂, a personal health environment for diabetic patients and the medical professionals that treat them. In figure 7 presents architecture showing how to extend the COMMODITY₁₂ PHS [30], with Ontologies and ORC agents to support semantic reasoning for diabetes patient profiles.

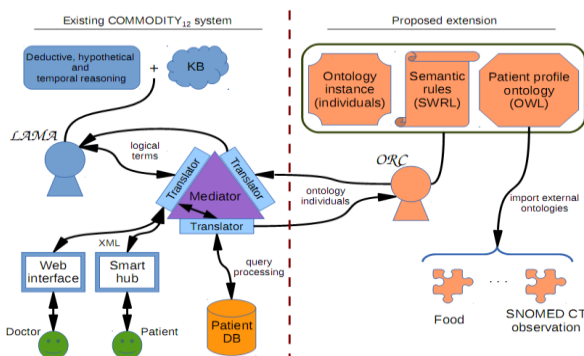


Fig.7. Architecture showing how to extend the COMMODITY12 PHS with Ontologies [29]

In [31], the authors developed a Diabetes Medication Recommendation system, based on domain Ontology that employ the knowledge base provided by a hospital specialist in Taichung’s Department of Health and the database of the American Association of Clinical Endocrinologists Medical Guidelines for Clinical Practice for the Management of Diabetes Mellitus (AACEMG). By thorough analysis, the system first builds ontology knowledge about the drugs’ nature attributes, type of dispensing and side effects, and ontology knowledge about patients’ symptoms. It then utilizes Semantic Web Rule Language (SWRL) and Java Expert System Shell (JESS) to induce potential prescriptions for the patients. This system is able to analyze the symptoms of diabetes as well as to select the most appropriate drug from related drugs.

In [32], the authors purposed is to offer through three simple stages a solution based on Ontologies to provide personalized monitoring services for patients with any of a wide range of chronic conditions in a tele-monitoring scenario. Presenting the work through the three stages, actions involved in each stage are clearly described enhancing its understanding, reusability and transferability of both the Ontology and the methodology for different domains or applications.

In [33], the authors developed Ontology for the care of chronically ill patients and implement two personalization

processes and a decision support tool. The first personalization process adapts the contents of the Ontology to the particularities observed in the health-care record of a given concrete patient, automatically providing a personalized Ontology containing only the clinical information that is relevant for health-care professionals to manage that patient. The second personalization process uses the personalized Ontology of a patient to automatically transform intervention plans describing health-care general treatments in to individual intervention plans. For comorbid patients, this process concludes with the semi-automatic integration of several individual plans in to a single personalized plan.

Finally, the ontology is also used as the knowledge base of a decision support tool that helps health-care professionals to detect anomalous circumstances such as wrong diagnoses, unobserved comorbidities, missing information, unobserved related diseases, or preventive actions.

VI. COMPARATIVE AMONG ONTOLOGY BASED SYSTEM FOR HEPATOBILIARY SYSTEM DISEASES

There are many research works described different types of diseases and Ontological engineering approach on Hepatobiliary system. These works have been achieved to build specific domain Ontologies and systems for different diseases in Hepatobiliary system as shown in table 1.

Table 1: Represents comparison among Ontology-based systems for Hepatobiliary System Diseases

System Name	Year	Scope	System							Protégé (OWL)	SWRL
			Diagnosis	Web Service	Disease causal chain	Treatment plan	Medicine side effect for disease	Query	Tele monitoring		
Browsing tool	2012	Diabetes	•	•	✓	•	•	•	•	•	•
VHOSWS	2012	Viral Hepatitis	✓	✓	•	•	•	✓	•	✓	•
Representation for Liver Cancer	2012	Liver Cancer	✓	•	•	•	•	•	•	✓	•
ORC	2013	Diabetes	•	•	•	•	•	•	•	✓	✓
Personalized monitoring services	2013	Chronic	•	•	•	•	•	•	✓	✓	•
SonaRes	2011	Gallbladder	✓	•	•	•	•	•	•	✓	✓
OUI	2014	hepato-pancreo-biliary	✓	•	•	•	•	•	•	✓	•
Liver Immunology Ontology (LIO)	2011	Liver Immune	•	•	•	•	•	✓	•	✓	•
Recommendation system for anti-diabetic drugs	2012	anti-diabetic drugs	✓	•	•	•	•	•	✓	•	✓
Health-care Ontology	2012	Chronic	✓	•	•	•	•	•	•	✓	•

Browsing tool is system for disease casual chain in diabetes. VHOSWS is web service system for viral hepatitis diagnosis and query by using Ontology which is built by protégé editor (OWL file). Representation for Liver Cancer is system for liver cancer diagnosis by using Ontology which is built by protégé editor (OWL file). Ontology reasoning component (ORC) is Ontology in diabetes field which is built by protégé editor (OWL file) using Semantic Web Rule

Language (SWRL). A personalized monitoring service is system for chronic tele- monitoring which is built by protégé editor (OWL file). While, SonaRes is system for Gallbladder diagnosis which is built by protégé editor (OWL file). Ontology of Ultrasound Images (OUI) is system for hepato-pancreato-biliary zone organs diagnosis which is built by protégé editor (OWL file). Liver Immunology Ontology (LIO) is system for LIO query which is built by protégé editor (OWL file). Recommendation system based for anti-diabetic drugs diagnosis and query which is built by protégé editor (OWL file) using SWRL. On the other hand, health-care Ontology is system for chronic diagnosis which is built by protégé editor (OWL file).

VII. CONCLUSION AND FUTURE WORK

This paper discussed the current studies on Ontological engineering approach and Ontology techniques for Hepatobiliary system diseases. It presented conceptual view of Ontology based Hepatobiliary system and its infected diseases. Furthermore, it presented a proposed a new classification schema for the research efforts investigated so far based on the Hepatobiliary system organs: Liver, Gallbladder, Bile duct and Pancreas. Beside, paper shows researchers worked in vary systems and Ontological engineering approach on Hepatobiliary system. The current work builds a new Ontology which using protégé editor (OWL file) and system for diagnosis, disease causal relation and query. Additionally, medicine side effects for disease, treatment plan and electronic patient records (EPRs) on different types of diseases in Hepatobiliary system.

REFERENCES

- [1] Cicortas, A., Iordan, V., Fortis, A., "Considerations on Construction Ontologies", Journal Annals Computer Science Series 1, 79–88, 2009.
- [2] W3C XML Schema Definition Language, <http://www.w3.org/TR/xmlschema11-1/>, last visited 12 Oct, 2014.
- [3] Resource Description Framework (RDF) Schema Specification. <http://www.w3.org/TR/PR-rdf-schema/>, last visited 29 Oct, 2014.
- [4] DAML+OIL Web Ontology Language. <http://www.w3.org/Submission/2001/12/>, last visited 29 Oct, 2014.
- [5] Web Ontology Language (OWL) Overview. <http://www.w3.org/TR/owl-features/>, last visited 29 Oct, 2004.
- [6] Zhanjun Li, Maria C. Yang and Karthik Ramani, "A methodology for engineering ontology Acquisition and validation", Artificial Intelligence for Engineering Design, Analysis and Manufacturing, PP: 37–51, USA, 2009.
- [7] Moise Gabriela, Netedu Loredana, "Ontologies for Interoperability in the eLearning Systems", Bulletin of Petroleum-Gas University of Ploiesti Mathematics, Informatics, Physics Series [BMIF], ISSN 1224-4899, EISSN 2067-242X, Volume LXI No. 2, PP: 75-88, 2009.
- [8] Youn Seongwook, Arora Anchit, Chandrasekhar Preetham, Jayanty Paavany, Mestry Ashish and Sethi Shikha. "Survey about Ontology Development Tools for Ontology-based Knowledge Management", University of Southern California, 2009.
- [9] H. Pundt and Y. Bishr. "Domain Ontologies for Data Sharing-An Example from Environmental Monitoring Using Field GIS", Computer & Geosciences, 28, PP: 98-102, 1999.
- [10] Protégé. <http://protege.stanford.edu/>, last visited 29 Oct, 2012.

- [11] Manolis Maragoudakis and Ilias Maglogiannis, "A medical ontology for intelligent web-based skin lesions image retrieval", Journal Health Informatics Journal, PP: 140-157, 2011.
- [12] Sánchez, D., Moreno, A. "Learning Medical Ontologies from the Web", LNCS Knowledge management for Health care Procedures, Vol. 4924, PP: 32-45, 2008.
- [13] Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector AL & Rosse C, "Relations in biomedical ontologies", Genome Biology, Vol. 6, 2005.
- [14] Samantha Bail, Matthew Horridge, Bijan Parsia, Ulrike Sattler, "The Justificatory Structure of the NCBO BioPortal Ontologies", Proceeding of The Semantic Web ISWC, Vol. 7031, PP: 67-82, 2011.
- [15] C. Paul Morrey, James Geller, Michael Halper & Yehoshua Perl, "The neighborhood auditing tool: A hybrid interface for auditing the UMLS", Journal of biomedical informatics, Vol. 42, PP: 468-489, 2009.
- [16] Mythili Thirugnanam, Mangayarkarasi Ramaiah, V Pattabiraman, R Sivakumar, "Ontology based Disease Information System", International Conference on Modelling Optimization and Computing, PP: 3235- 3241, 2012.
- [17] Official Partner of the Liver Strong Foundation, <http://www.livestrong.com/article/119869-list-body-systems>, last visited 10 December, 2014.
- [18] Medicine Net, http://www.medicinenet.com/liver_disease/article.htm/, last visited 10 December, 2014.
- [19] <http://hepatitis.about.com/od/jkl/g/liver.htm/>, last visited 10 December, 2014.
- [20] Moawad,I, Al Marzoqi,G, Salem, A," Web Service Based Approach for Viral Hepatitis Ontology Sharing and Diagnosing", In Proceeding of AMLTA, PP: 257-266, Springer-Verlag Berlin Heidelberg 2012.
- [21] Marco Alfonse, Mostafa M. Aref, Abdel-Badeeh M. Salem. "Ontology-Based Knowledge Representation for Liver Cancer". Proceedings of the International eHealth, Telemedicine and Health ICT Forum for Educational, Networking and Business. Luxembourg, G. D. of Luxembourg, ISSN 1818 -9334, PP: 821-825, April 18-20, 2012.
- [22] Anna Maria Masci, Jeffrey Roach, Bernard de Bono, Pierre Grenon, Lindsay Cowell, "Bridging multiple Ontologies Representation of the liver Immune Response", International Conference on Biomedical Ontologies (ICBO), Buffalo,NY, USA, Working with multiple Biomedical Ontologies Workshop, 2011.
- [23] NatalieBruc, GalinaMagariu, TatianaVerlan," Gallbladder Description in Ultrasound Images Ontology", Proceeding of Modelling and Development of Intelligent Systems, PP: 18-27, Sibiu-Romania, 2011.
- [24] Natalia BRUC," An approach to mapping between the classification of diseases ICD-10 and the Ontology of Ultrasound Images of hepato-pancreato-biliary zone organs", 8th International Conference on Microelectronics and Computer Science, Chisinau, Republic of Moldova, PP: 314-317, 2014.
- [25] DICOM, Available: <http://dicom.nema.org/>, last visited 12 January, 2015.
- [26] International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10), Available: <http://apps.who.int/classifications/icd10/browse/2010/en#K80-K87>, last visited 4 December, 2014.
- [27] N. Bruc, G. Magariu, T. Verlan, "Elaborating of Ultrasound Images Ontology in Ultrasound Diagnostics", in the International Conference on e-Health and Bioengineering, EHB, Iasi, In CD, Gr.T.Popa University of Medicine and Pharmacy Publishing House, Iași, România, Editors: Hariton Costin, Alexandru Morega, Liliana Vereștiuc. ISBN: 978-606-544-078-4, 2011.
- [28] Kouji Kozaki, Hiroko Kou, Yuki Yamagata, Takeshi Imai, Kazuhiko Ohe, Riichiro Mizoguchi," Browsing Casual Chain in a Disease Ontology", International Semantic Web Conference (Posters & Demos)'12, 2012.

- [29] Kafali, Ozgur; Sindlar, Michal; Weide, Tom van der; Stathis, Kostas,” ORC: an Ontology Reasoning Component for Diabetes”, 2nd International Workshop on Artificial Intelligence and Netmedicine (NetMed’13). 2013.
- [30] Kafali,Ö.,Bromuri,S.,Sindlar,M.,vanderWeide,T.,Pelaez,E.A.,Schaechtle,U.,Stathis,K, “COMMODITY12:A smarte-health environment for diabetes management”, Journal of Ambient Intelligence and Smart Environments, IOS Press (Toappear), 2013.
- [31] Rung-Ching Chen, Yun-Hou Huang, Cho-Tsan Bau, Shyi-Ming Chen,” A recommendation system based on domain ontology and SWRL for anti-diabetic drugs selection”, Expert Systems with Applications, Volume 39, Issue 4, PP: 3995-4006, 2012.
- [32] Lasiera N , Alesanco A, Guillén S, García J, “A three stage ontology-driven solution to provide personalized care to chronic Patients at home”, Journal of Biomedical Informatics, PP: 516-529, 2013.
- [33] David Riaño, Francis , Joan Albert López-Vallverdú, Fabio Campana, Sara Ercolani, Patrizia Mecocci,, Roberta Annicchiarico, Carlo Caltagirone,” An ontology-based personalization of health-care knowledge to support clinical decisions for chronically ill patients”, Journal of Biomedical Informatics Volume 45, Issue 3, PP: 429-446, 2012.

Hierarchical Sparsity-Regularized Framework Based Frequency Hopping Spectrum Estimation With Antenna Array System

Lifan Zhao, Lu Wang, Bi Guoan

School of Electrical and Electronic Engineering
Nanyang Technological University
Singapore
{zhao0145,wa0001lu,egbi}@ntu.edu.sg

Haijian Zhang

School of Electronic Information
Wuhan University
Wuhan, China
haijian.zhang@whu.edu.cn

Abstract— Frequency hopping (FH) signals have been widely employed in wireless communication networks to combat interference and avoid collision. This paper considers the blind FH signal estimation problem in antenna array systems, where the direction-of-arrivals, hopping time and frequency are all unknown to the users. A hierarchical sparsity-aware technique is developed to estimate these parameters in an optimization framework. More concretely, sparsity in spatial domain and time-frequency domain are exploited in a hierarchical and iterative manner, respectively, where more accurate parameter estimation performance can be obtained. Compared to prior state-of-the-arts, conventional model-order selection procedure can be conveniently avoided due to the utilization of sparsity-regularized framework. Results of numerical experiments show that the proposed algorithm can achieve superior performance particularly in sub-Nyquist sampling and low signal-to-noise ratio (SNR) scenarios compared with other recently reported ones.

Keywords—Frequency hopping, antenna array systems, sparsity-regularized framework

I. INTRODUCTION

In wireless communications and other information systems, frequency hopping (FH) signals have been widely employed due to their advantageous capability of low interception and anti-jamming [1]–[3]. In many wireless networks such as home area networks (HAN) and Blue tooth personal area network(PAN), FH signal is employed to combat near-far problem[4] and avoid collision [1], [5], [6]. When non-cooperative FH networks coexist, however, it is inevitable that the receiver encounters multiple unknown FH signals from different emitters. In these systems, antenna array is often utilized to provide spatial degree of freedom to separate sources from different directions. The main challenge is to robustly estimate the DOAs, hopping time and frequency of the multiple FH signals. Since the maximum likelihood estimation will induce intractable computations and cause over-fitting problems [7],[8], many other alternative approaches are developed to achieve robust estimation.

The conventional time-frequency distribution (TFD) is a natural tool to exhibit the non-stationary frequency content

of the signals. DOA estimation procedure, such as beamforming or multiple signal classification (MUSIC), is often carried out before applying TFD to estimate the hopping time and frequency in a non-parametric manner. The linear TF distribution such as, short-time Fourier transform (STFT), is a popular choice for estimating the parameters of each FH signal due to it is free of cross-terms [9], while quadratic TFD such as, Wigner-Ville distribution (WVD) is also employed for its high energy concentration of the signal [10]. Even though entropy or gradient based refinement is often employed to the TFD as a post-processing procedure, either limited resolution (linear TFD) or undesirable cross-terms (quadratic TFD) greatly inhibit substantial improvements by these procedures. In [1], [11], another method is developed to cope with DOA, hopping time and frequency estimation in FH networks. An expectation maximization (EM) approach is employed to iteratively estimate the amplitude, hopping time and frequency. In this approach, the success of the EM algorithm depends largely on the proper selection of initial value, where a denoised STFT is often used to initialize the algorithm. In low SNR environments, however, it is hard to obtain a desirable initial TFR, which will inevitably result in degraded

performance of the EM algorithm. More importantly, the model order selection procedure is required to facilitate the estimation. To effectively deal with the heavy burden on A/D hardware, multiple FH signal estimation is considered in a sparse linear regression (SLR) framework [8], [12], where a fused-lasso like formulation is employed to solve this problem. In this framework, two penalty terms are defined to encourage sparsity [13] and smoothness in the TF domain, respectively. The formulation can be expressed as,

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x}} \left[\frac{1}{2} \|\mathbf{y} - \mathbf{W}\mathbf{x}\|_2^2 + \lambda_1 \|\mathbf{x}\|_1 + \lambda_2 \|\mathbf{D}\mathbf{x}\|_1 \right] \quad (1)$$

where the first term is for data fitting, the second and third terms are for encouraging sparsity in frequency and differential-time domain, respectively [8]. One remarkable advantage of this approach is that it does not require to select the order of the parameters. Moreover, due to the grid based formulation, this approach can even be conveniently accommodate to different sampling strategy [8]. However, this approach is particularly formulated for single channel system, where non-trivial modification is required to be carried out for multi-channel systems. Inspired by the SLR framework, we propose a hierarchical sparsity-regularized framework to cope with the abovementioned challenges. In the first stage of the framework, spatial sparsity of the signals is exploited, while frequency and hopping sparsity are exploited in the second stage of the framework. The main idea of the algorithm is to properly utilize the hierarchical sparsity model to achieve jointly sparsity. By unifying these two stages in an iterative estimation, more accurate estimation performances can be expected. The proposed method can be considered as a model-selection free approach, where model-order mismatch problem is desirably avoided by the utilization of sparsity. Moreover, the proposed approach can allow for mutual improvements of parameter estimation due to the proposed iterative scheme. The paper is organized as follows. In Section II, the preliminary is presented and the hierarchical sparse model is introduced. In Section III, the sparsity-regularized framework is formulated in a hierarchical manner. The proposed approach is proposed in Section IV of the paper. Finally, the experimental results are given in Section V to validate the performance of the proposed algorithms.

Notation : Vectors and matrices are denoted by bold symbol. For a vector \mathbf{x} , \mathbf{x}^* and \mathbf{x}^H represent the conjugate and Hermite of the vector. For a matrix \mathbf{A} , \mathbf{A}^H and \mathbf{A}^{-1} denote the conjugate and inverse of the matrix, respectively. l_p norm of the vector or matrix is denoted by $\|\cdot\|_p$. $CN(\mathbf{X}/\mu; \Sigma)$ denotes multivariate complex Gaussian distributions.

II. PRELIMINARY AND PROBLEM FORLUMATION

In the multi-channel FH sensor network, the direction-of arrival (DOA), hopping time and frequency are all required to be estimated by each user. In particular, the geometry of the uniform linear array (ULA) is given in Fig. 1. In this section, we present the received signal model of the multiple FH

signals. Assuming K_s sources are impinging on a ULA with L sensors, the received signal in each snapshot n is given as

$$\mathbf{y}(n) = \Phi \mathbf{s}(n) + \mathbf{v}(n), n = 1, \dots, N_s \quad (2)$$

where $\mathbf{y}(n) \in C^{L \times 1}$, $\Phi = [\mathbf{a}(\theta_1); \dots, \mathbf{a}(\theta_{N_0})] \in C^{L \times N_0}$, $\mathbf{s}(n) \in C^{N_0 \times 1}$ contains the FH signals, N_s is the number of snapshots. The noise $\mathbf{v}(n)$ is assumed to be spatially and temporally uncorrelated,

$$\mathbb{E}[\mathbf{v}(n_1)\mathbf{v}^H(n_2)] = \alpha_0^{-1} \mathbf{I} \cdot \delta_{n_1, n_2}. \quad (3)$$

Consider a FH network where each user receives the noise corrupted multiple FH signals. Sampling the continuous signal $s(t)$, we can obtain the discrete form of $s(t)$ expressed as

$$s(n) = \sum_{k=1}^{K_i} \sigma_{i,k} e^{j2\pi f_{i,k} n / f_s}, \quad n_{i-1} < n < n_i \quad (4)$$

where K_i is the number of hopping frequencies during the i th system dwell time, $\sigma_{i,k}$ is the amplitude of the signal associated with the i th system dwell time and k -th hopping frequency. Since the noise obeys complex Gaussian distribution, the likelihood of the received signal can be expressed as

$$p(\mathbf{Y}|\mathbf{S}) = \prod_{k=1}^{N_s} \mathcal{CN}(\mathbf{y}(k) | \Phi \mathbf{s}(k), \sigma^2) \quad (5)$$

In practical communication systems, direction of arrivals, the system dwell time and hopping frequency are all unknown to the user, and required to be estimated robustly to avoid retransmission. To illustrate the difficulty of the problem, the maximum likelihood (ML) estimation can be formulated as the following optimization problem given as,

$$(\hat{\mathbf{n}}_k, \hat{K}, \hat{f}, \hat{N}_k, \hat{\sigma}) = \arg \min_{\hat{\mathbf{n}}_k, \hat{K}, \hat{f}, \hat{N}_k, \hat{\sigma}} \sum_{k=1}^{\hat{K}} \left\| \mathbf{Y}_k - \sum_{n=1}^{\hat{N}_k} \hat{\sigma}_{kn} B(\hat{f}_{kn}) \right\|_F^2 \quad (6)$$

In blind frequency hopping signal estimation problem dealt within this paper, the frequency location, hopping time instant, the number of components are all unknown, which is a very challenging problem [8]. In conventional estimation scheme, model order should be selected for good performances. The above formulation is known to be a non-convex optimization problem and is analytically non-solvable [11], which is equivalent to the ML estimation. Since the hopping instant n_k , the number of hopping instants \hat{K} , the hopping frequency $f_{i,k}$, the number of components N , complex magnitude σ_{kn} and even direction of arrivals θ_k (in multi-channel case) are all unknown, solving this problem by exhaustive search is computational intractable. It is argued that the ML estimation by searching are these parameters in a combinational way is not a trivial task [1], [8]. Moreover, proper model-order selection procedure is also required to be carried out to avoid over-fitting problem.

III. HIERACHICAL SPARSITY-REGULARIZED FRAMEWORK

Inspired by the sparsity regularized framework, in this section, a hierarchical framework is introduced to estimate the

DOA and FH signal in a sparsity-driven manner. As presented in Section II, the estimation problem can be expressed in a hierarchical linear model. In sensor array system, the sparsity of the signals in spatial domain and time-frequency domain can be exploited, respectively. More concretely, the signals are only coming from a small number of angles, and can be represented in the time-frequency domain in a piecewise sparse manner. Remarkably, this sparse driven approach can

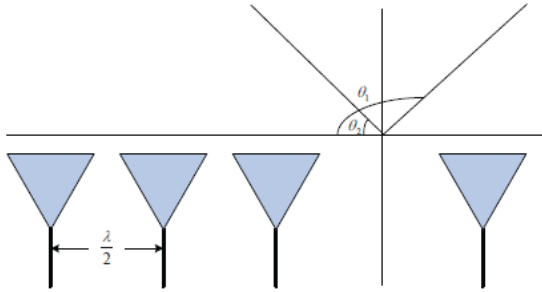


Fig. 1. The geometry of the ULA with two sources coming from θ_1 and θ_2 . The antenna interval is set as $\lambda/2$.

obtain desirable estimation results and avoid the model-order selection procedure.

A. Stage 1: Spatial Sparsity

To estimate the DOA with sparsity constraint [14], [15], a new matrix $\tilde{\mathbf{A}} \triangleq [\mathbf{a}(\theta_1), \mathbf{a}(\theta_2), \dots, \mathbf{a}(\theta_{N_0})] \in \mathbb{C}^{L \times N_0}$ is particularly manipulated as an over-complete dictionary, whose i -th atom

$$\mathbf{a}(\theta_i) = [e^{j\pi \cos \theta_i}, e^{j2\pi \cos \theta_i}, \dots, e^{jL\pi \cos \theta_i}]^H \quad (7)$$

corresponds to an ideal steering vector from spatial angle θ_i , based on the geometry of the ULA presented in Fig. 1. Based on the sparse representation, the signal model can be constructed as,

$$\mathbf{Y} = \tilde{\mathbf{A}}\mathbf{S} \quad (8)$$

where \mathbf{S} is a row-sparse matrix. The non-zero rows of the matrix \mathbf{S} represent the FH signals from different directions. In this stage, the prior for the row-sparse matrix \mathbf{S} can be expressed as,

$$p(\mathbf{S}) \propto \exp \left(-\lambda \sum_{k=1}^{N_0} \left| \sqrt{\sum_{i=1}^N x_{ki}^2} \right| \right) \quad (9)$$

where l is defined as a hyper-parameter to control the rowwise sparsity degree of the matrix \mathbf{S} . In particular, a non-zero row of the matrix \mathbf{S} corresponds a FH signal from a particular direction.

B. Stage 2: Hopping Time And Frequency Sparsity

Apart from the spatial sparsity of the received signal, the FH signal exhibits hopping time and frequency sparsity in time-frequency domain in the modeling of the second stage

[16]. Let us denote the non-zero row of \mathbf{S} as $\sim \mathbf{S}_i$. To exploit the hopping time and frequency sparsity, the signal model can be represented by,

$$\mathbf{S}_i^T = \mathbf{W}\mathbf{x}_i. \quad (10)$$

where $\mathbf{x}_i \in \mathbb{C}^{N_s \times 1}$ is the stacked time-frequency coefficients for the i -th FH signal and $\mathbf{S}_i \in \mathbb{C}^{1 \times N_s}$ is the FH signal from the i -th direction. The i -th row of matrix \mathbf{W} is constructed as $\mathbf{W}_i = [0_N^H, \dots, 0_N^H, \boldsymbol{\omega}^T, 0_N^H, \dots, 0_N^H]^T$ where $\boldsymbol{\omega} = [e^{j\omega_1 n}, \dots, e^{j\omega_N n}]^T$ is a vector representing frequency grid. To exploit the frequency sparsity, the prior for \mathbf{x}_i is also a multi-variate Laplace distribution, expressed as [17]

$$p_1(\mathbf{x}_i) \propto \exp(-\lambda_1 \|\mathbf{x}_i\|_1) \quad (11)$$

where the matrix \mathbf{D} is a differential matrix defined in [8], [12]. Combining these two priors will result in both the hopping time and frequency prior for \mathbf{x}_i .

Based on the above described can be expressed as maximum a posterior (MAP) estimation, the solution of can be expressed as

$$(\mathbf{S}, \mathbf{x}) = \arg \max_{\mathbf{S}, \mathbf{x}} p(\mathbf{Y}|\mathbf{S})p(\mathbf{S})p_1(\mathbf{x})p_2(\mathbf{x}) \quad (13)$$

s.t. $\mathbf{S}_i^T = \mathbf{W}\mathbf{x}_i$

However, this MAP estimation problem is a hard to be solved directly due to the explicit coupling of prior for \mathbf{S} and \mathbf{x} . Therefore, we propose an approximate algorithm to solve this problem in Section IV.

Remark 1: The above-describe model will naturally incorporate the sparsity in spatial, hopping time and frequency domain in a hierarchical manner. However, how to properly and effectively exploit this hierarchical sparsity to obtain mutual enhancement of DOA and FH signal estimation is the key issue. More concretely, exploiting the sparsity in the spatial domain should enhance the FH signal estimation performance and vice versa.

IV. FH SIGNAL ESTIMATION PROBLEM

In this section, a hierarchical sparse regularized estimation (HSRE) approach is developed to obtain DOA and FH signal estimation in a joint manner. By leveraging the hierarchical sparsity, the immediate advantage of the proposed approach is that it can desirably avoid model-order selection procedure. In particular, an iterative procedure is proposed to estimate the DOA and FH signals in a hierarchical way. Another remarkable advantage of this approach is that it can achieve much better results by iteratively refining the DOA estimation and FH signal estimation. In this approach, the DOA estimation and FH signal estimation is carried out in separate stages, while allowing mutual improvement of the parameter estimation accuracy.

DOA estimation Stage: Assume that the FH signal has been estimated from the last iteration. To properly utilize the information, we propose to carry our the beam-forming procedure for the i -th FH signal, which can be expressed as,

$$\begin{aligned} \tilde{\mathbf{Y}}_1^i &= \mathbf{Y}\mathbf{S}_i^H = \left(\sum_{k=1}^K \mathbf{a}(\theta_k)\mathbf{S}_k + \mathbf{V} \right) \cdot \mathbf{S}_i^H \\ &= \|\mathbf{S}_i\|_2^2 \mathbf{a}(\theta_i) + \sum_{k=1, k \neq i}^K \mathbf{a}(\theta_k)\mathbf{S}_k\mathbf{S}_i^H + \mathbf{V} \cdot \mathbf{S}_i^H \quad (14) \end{aligned}$$

Since the correlation between different FH signals is relatively low, the second and third term in (14) are considered as new noise term. Therefore, in this stage, the modified likelihood function of the filtered signal can be represented by

$$p(\tilde{\mathbf{Y}}_1^i | \tilde{\mathbf{s}}_i) \sim \mathcal{CN}(\mathbf{Y}\mathbf{S}_i^H | \mathbf{A}\tilde{\mathbf{s}}_i, \sigma_1^2) \quad (15)$$

where $\tilde{\mathbf{s}}_i$ is a sparse vector denoting the spatial sparsity of the i -th FH signal. Combining the sparsity prior of $\tilde{\mathbf{s}}_i$, the MAP estimation of this problem can be given as,

$$\tilde{\mathbf{s}}_i = \arg \min_{\tilde{\mathbf{s}}} \left\| \mathbf{Y}\mathbf{S}_i^H - \mathbf{A}\tilde{\mathbf{s}}_i \right\| + \lambda \|\tilde{\mathbf{s}}_i\|_1 \quad (16)$$

where λ is the regularization parameter to balance data fitting and sparsity.

Remark 2: Based on the estimation of the FH signal, the signal is projected on the estimated FH signal, respectively. The underlying rationale is that the DOA estimation can be more accurately carried out based on the projection. The DOA can be more accurately estimated, since the filtered signal is a one-sparse signal in spatial domain. FH signal Estimation Stage: Assume that the DOAs have been estimated in the above stage. Similarly, the beam-forming procedure is carried out to estimate the FH signal for the i -th direction as,

$$\begin{aligned} \tilde{\mathbf{Y}}_2^i &= \mathbf{a}(\hat{\theta}_i)^H \mathbf{Y} = \mathbf{a}^H(\hat{\theta}_i)\mathbf{A}\mathbf{s} + \mathbf{a}^H(\hat{\theta}_i)\mathbf{V} \\ &= \mathbf{s}_i + \sum_{j=1, j \neq i}^{K_s} \mathbf{a}^H(\hat{\theta}_i) \cdot \mathbf{a}(\theta_j) \cdot \mathbf{s}_j + \mathbf{a}^H(\hat{\theta}_i)\mathbf{V}. \quad (17) \end{aligned}$$

Therefore, the modified likelihood function of the beamformed signal can be represented by,

$$p(\tilde{\mathbf{Y}}_2^i | \mathbf{x}_i) \sim \mathcal{CN}(\mathbf{Y}^H \mathbf{a}(\hat{\theta}_i) | \mathbf{W}\mathbf{x}_i, \sigma_2^2) \quad (18)$$

where \mathbf{x}_i is a sparse vector denoting the time-frequency coefficient of the FH signal from i -th direction. Combining the sparsity prior of $\tilde{\mathbf{x}}_i$ in both hopping time and frequency domain, the MAP estimation of this problem can be given as,

$$\mathbf{x}_i = \arg \min_{\mathbf{x}_i} \left\| \mathbf{Y}^H \mathbf{a}(\hat{\theta}_i) - \mathbf{W}\mathbf{x}_i \right\| + \lambda_1 \|\mathbf{x}_i\|_1 + \lambda_2 \|\mathbf{D}\mathbf{x}_i\|_1 \quad (19)$$

Remark 3: Based on the estimation of the DOA, beamforming procedure is carried out before FH signal estimation. An immediate advantage is that this procedure will separate the FH signals, where each one can be considered to be sparser in both hopping time and frequency domain. Therefore, the proposed algorithm operates in an iterative manner as shown in Algorithm 1. It is noted that the number of sources K_s is not required to be estimated by model-selection technique, rather it can be easily determined by thresholding residual due to sparsity, which is not discussed further for brevity.

V. FH SIGNAL ESTIMATION PROBLEM

In this section, the experimental results are presented to validate the performance of the algorithm. The regularization parameter for FH signal estimation stage is chosen to be $\lambda_1 = \lambda_1^*/10$ and $\lambda_2 = \lambda_2^*/10$ to achieve desirable hopping time and frequency detection in the tested SNRs, where optimal λ_1^* and λ_2^* are given in [8]. The number of Monte Carlo trails is chosen to be 100.

The hopping signals used in the following experiments are generated as follows: the first hopping component is active within the range of time index [0 : 15] and the carrier frequency hops from 13 KHz to 18 KHz within the range of time index [16 : 63]. The second hopping component is active within the range of time index [0 : 31] and the carrier frequency hops from 28 KHz to 23 KHz within the range of time index [32 : 63].

Algorithm 1 Multi-channel FH signal estimation

```

1: Input:  $\mathbf{Y}, \Phi$ 
2: % Initialization: overcomplete dictionary  $\mathbf{A}$ , initial DOA estimation, differential dictionary  $\mathbf{D}$ ;
3: while ~Converge do
4:   % I. DOA Estimation Stage;
5:   for  $j = 1 : K_s$  do
6:     Obtain filtered signal  $\mathbf{Y}\mathbf{S}_j^H$ ;
7:     Use  $l_1$ -minimization or OMP [19] algorithm to estimate DOA;
8:   end for
9:   % II. FH Signal Estimation Stage
10:  for  $j = 1 : K_s$  do
11:    Beamforming  $\tilde{\mathbf{Y}}_j = \mathbf{a}^H(\hat{\theta}_j)\mathbf{Y}$ 
12:    Estimate  $\mathbf{x}_j$  from (19)
13:  end for
14: end while
15: Output:  $\mathbf{x}_i, i = 1, \dots, M$ .
```

In the following, the frequency hopping signal estimation based on EM (FHSE-EM) algorithm is particularly compared for performance evaluation. In the following experiments, the SNR is defined as

$$\text{SNR} = 10 \log_{10} \left(\frac{\|\mathbf{s}\|_2^2}{N_s \sigma^2} \right) \quad (20)$$

where \mathbf{s} denotes the signal vector, N_s is the number of time indices and σ^2 is the noise power.

In particular, two performance measures are defined for comparison of hopping time and instantaneous frequency (IF) detection, respectively. The correct hopping time detection ratio is defined as,

$$P_t = \left(\sum_{i=1}^{M_c} D_t(i) \right) / M_c \quad (21)$$

where M_c is the number of Monte Carlo trails and $D_t(i)$ is the number of correct detections in each Monte Carlo trial. The hopping time statistic is defined as $\Delta_n = \|x_{n+1} - x_n\|_2^2$. A correct hopping time detection is declared if the estimated hopping instant is less than 3 samples away from the associated true hopping instant, which is defined in the same

way as in [8], [18]. The incorrect IF detection ratio is further defined as,

$$E_f = 10 \log_{10} \left(1 - \left(\sum_{i=1}^{M_c} D_f(i) \right) / M_c \right) \quad (22)$$

where $D_f(i)$ is the correct frequency detection rate in the i -th Monte Carlo trial. An illustrative example is given in Fig. 2. In this experiment, the true DOA of the two sources are $\theta = [40; 60]$. It is seen that the spectrogram obtained in Fig. 2 (a) and (c) cannot give a good energy concentration of the signal, particularly in the first system dwell time. In contrast, the time frequency representation obtained by proposed sparsity-driven methods in Fig. 2 (e) and (g) can give much better estimation and concentration due to the utilization of hierarchical sparsity regularized strategy in the time-frequency domain and spatial domain, respectively. Notably, the time-frequency representation obtained by the proposed method is considered quite a desirable one since the noise is well pruned away with clear hopping time estimation compared to those obtained by

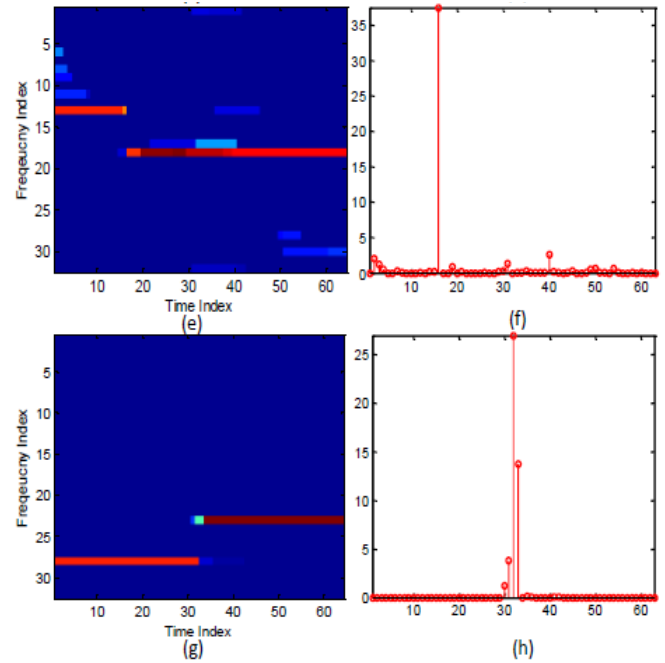


Fig. 2. An illustrative example of frequency hopping signal estimation. (a) Spectrogram (component 1) and (b) the hopping time statistic; (c) Spectrogram (component 2) and (d) the hopping time statistic; (e) proposed HSRE (component 1) and (f) the hopping time statistic; (g) proposed HSRE (component 2) and (h) the hopping time statistic. (The DOA estimation for STFT is $(\theta = [38, 61])$ using ℓ_1 -svd and DOA estimated by HSRE is $(\theta = [40, 60])$)

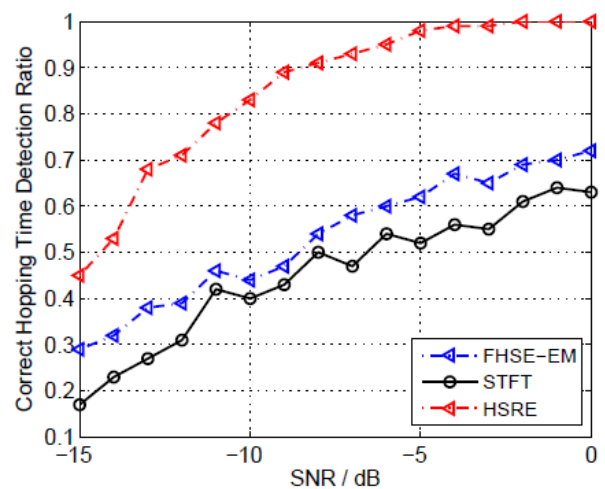
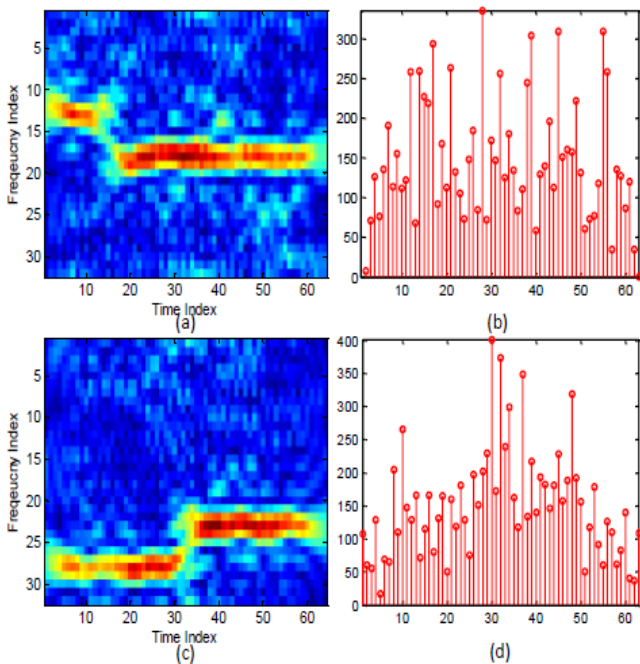


Fig. 3. The correct hopping time detection ratio for frequency hopping signal with two components is presented, in terms of SNR. The performance of the proposed approach is compared with FHSE-EM and STFT. The number of sensors is set to be 8.

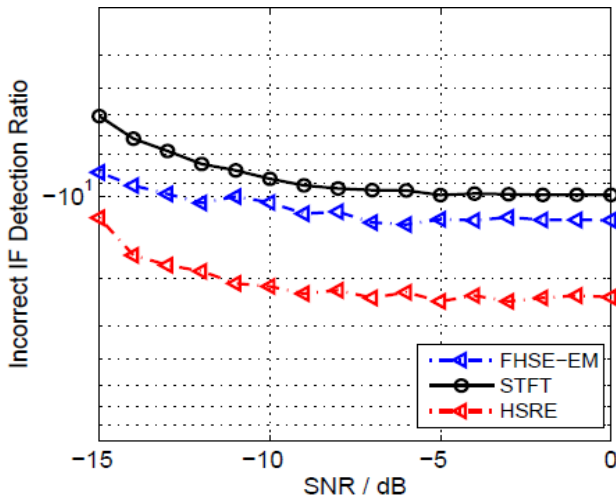


Fig. 4. The incorrect IF detection ratio for frequency hopping signal with two components is presented. The number of sensors is set to be 8.

STFT. The corresponding hopping time statistics also demonstrate the superiority of the proposed method. As seen from Fig. 2 (b) and (d), the statistics does not give the correct and has undesirable side-lobes. In contrast, the hopping time statistics obtained by proposed algorithm, as shown in Fig. 2 (f) and (h) can give the correct estimation result and has a sharp spike in hopping time instant. Therefore, it can show that the proposed approach can achieve better DOA estimation accuracy and higher hopping time detection rate.

In Figs. 3, Monte Carlo experiments are conducted to give quantitative evaluation of the proposed algorithm and the previously reported ones, in terms of correct hopping time detection ratio. From this figure, it can be observed that with the increase of SNR, the STFT, FHSE-EM and the proposed all achieve better performance. In particular, the correct hopping time detection ratio of proposed HSRE method is almost 100%, when SNR > -5dB. Across all the SNRs, our proposed algorithm can achieve the best hopping time detection ratios due to the utilization of hierarchical sparsity.

In Fig. 4, the corresponding incorrect IF detection ratio is presented with same experimental settings. From this figure, similarly, all of these algorithms gives lower incorrect IF detection ratio with the increase of SNRs. In particular, our proposed algorithm can achieve the lowest incorrect IF detection ratio due to the inherent scheme of joint sparsity inducing procedure.

VI. CONCLUSION

In this paper, a new frequency hopping signal estimation method in sensor array system is developed, where DOA, hopping time and frequency can be jointly estimated. Compared with existing research, in our work, a hierarchical framework is particularly utilized to combat the difficulties of model order selection procedure, by exploiting spatial sparsity and time-frequency sparsity. This novel approach can not only avoid the tedious parameter tuning process, but also provide robust performance in low SNR environments.

REFERENCES

- [1] X. Liu, N. D. Sidiropoulos, and A. Swami, "Joint hop timing and frequency estimation for collision resolution in FH networks," *IEEE Transactions on Wireless Communications*, vol. 4, no. 6, pp. 3063–3074, 2005.
- [2] C. Chen and P. Vaidyanathan, "MIMO radar ambiguity properties and optimization using frequency-hopping waveforms," *IEEE Transactions on Signal Processing*, vol. 56, no. 12, pp. 5926–5936, 2008.
- [3] S. Srinivasa and S. A. Jafar, "Cognitive radios for dynamic spectrum access—the throughput potential of cognitive radio: A theoretical perspective," *IEEE Communications Magazine*, vol. 45, no. 5, pp. 73–79, 2007.
- [4] D. J. Torrieri, "Mobile frequency-hopping CDMA systems," *IEEE Transactions on Communications*, vol. 48, no. 8, pp. 1318–1327, 2000.
- [5] L. Zhang, H. Wang, and T. Li, "Anti-jamming message-driven frequency hopping - part I: System design," *IEEE Transactions on Wireless Communications*, vol. 12, no. 1, pp. 70–79, 2013.
- [6] L. Zhang and T. Li, "Anti-jamming message-driven frequency hopping - part II: Capacity analysis under disguised jamming," *IEEE Transactions on Wireless Communications*, vol. 12, no. 1, pp. 80–88, 2013.
- [7] X. Liu, Nicholas D. Sidiropoulos, and A. Swami, "Blind high-resolution localization and tracking of multiple frequency hopped signals," *IEEE Transactions on Signal Processing*, vol. 50, no. 4, pp. 889–901, 2002.
- [8] D. Angelosante, G.B. Giannakis, and N.D. Sidiropoulos, "Estimating multiple frequency-hopping signal parameters via sparse linear regression," *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5044–5056, 2010.
- [9] L. Cohen, "Time-frequency distributions—a review," *Proceedings of the IEEE*, vol. 77, no. 7, pp. 941–981, 1989.
- [10] S. Barbarossa and A. Scaglione, "Parameter estimation of spread spectrum frequency-hopping signals using time-frequency distributions," in *First IEEE Signal Processing Workshop on Signal Processing Advances in Wireless Communications*, 1997, pp. 213–216.
- [11] X. Liu, J. Li, and X. Ma, "An EM algorithm for blind hop timing estimation of multiple FH signals using an array system with bandwidth mismatch," *IEEE Transactions on Vehicular Technology*, vol. 56, no. 5, pp. 2545–2554, 2007.
- [12] D. Angelosante, G.B. Giannakis, and N.D. Sidiropoulos, "Sparse parametric models for robust nonstationary signal analysis: Leveraging the power of sparse regression," *IEEE Signal Processing Magazine*, vol. 30, no. 6, pp. 64–73, 2013.
- [13] S. S. Chen, D. L. Donoho, and M.A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Review*, vol. 43, no. 1, pp. 129–159, Jan. 2001.
- [14] D. Malioutov, M. C. etin, and A. S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Transactions on Signal Processing*, vol. 53, no. 8, pp. 3010–3022, 2005.
- [15] Z. Liu, Z. Huang, and Y. Zhou, "Sparsity-inducing direction finding for narrowband and wideband signals based on array covariance vectors," *IEEE Transactions on Wireless Communications*, vol. 12, no. 8, pp. 1–12, August 2013.
- [16] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused LASSO," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 1, pp. 91–108, 2005.
- [17] L. Zhao, G. Bi, L. Wang, and H. Zhang, "An improved auto-calibration algorithm based on sparse Bayesian learning framework," *IEEE Signal Processing Letters*, vol. 20, no. 9, pp. 889–892, 2013.
- [18] L. Zhao, L. Wang, G. Bi, L. Zhang, and H. Zhang, "Robust frequencyhopping spectrum estimation based on sparse bayesian method," to appear in *IEEE Transactions on Wireless Communications*.
- [19] E. J. Candes and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

Unsupervised Single Channel Source Separation with Nonnegative Matrix Factorization

A.M. Darsono, Shakir Saat, N.M. Z. Hashim, A.A.M ISA

Faculty of Electronics & Computer Engineering, Universiti Teknikal Malaysia Melaka,
 Melaka, Malaysia
 {abdmajid, shakir, nikzarifie, azmiawang}@utem.edu.my

Abstract— In this paper, a novel single channel source separation using two-dimensional nonnegative matrix factorization (NMF2D) is proposed. In NMF2D, the time-frequency (TF) profile of each source is modeled as two-dimensional convolution of the temporal code and the spectral basis. The proposed model used Beta-divergence as a cost function and updated by maximizing the joint probability of the mixing spectral basis and temporal codes using the multiplicative update rules. Results have concretely shown the effectiveness of the algorithm in blindly separating the audio sources from single channel mixture.

Keywords- Blind Source Separation; Nonnegative Matrix Factorization ; Machine Learning; Beta Divergence.

I. INTRODUCTION

Blind source separation (BSS) refers to the statistical technique of separating a mixture of underlying source signals. BSS has become one of the promising and exciting topics with solid theoretical foundations and potential applications in the fields of signal processing, neural computation and advanced statistics. Single channel source separation (SCSS) is a branch of BSS family where the blind signal separation is achieved when only one single recording is available. For many practical applications such as audio scenarios, generally only one channel recording is available in the hardware and in such cases conventional source separation techniques are not appropriate. Several approaches have been developed to solve the MSS problem such as the computer auditory scene analysis (CASA) [1] and underdetermined BSS [2, 3]. However both techniques are supervised technique which relies on a priori knowledge of sources obtained during the training phase to perform the separation. To overcome this, nonnegative matrix factorization (NMF) [4] approach is introduced where separation is performing without using any prior knowledge about the corresponding source signal. In NMF, given the matrix, \mathbf{Y} of a dimension of $F \times N$ with nonnegative elements, nonnegative matrix factorization (NMF) is the problem of approximate the factorization

$$\mathbf{Y} \approx \mathbf{W}\mathbf{H} \quad (1)$$

where $\mathbf{W} \in \mathfrak{R}^{F \times C}$ and $\mathbf{H} \in \mathfrak{R}^{C \times N}$ are a non-negative matrices. F represents the frequency bins while N represents the time slot in the TF domain. \mathbf{W} contains the spectral basis vectors while \mathbf{H} represents the amplitude of each basis vector at each time point. C is the numbers of component from data sources being used and it is determine such that $FC+CN \ll FN$ so that the data can be compressed to its integral component. This problem can be formulated as the minimization of an objective function.

$$D(\mathbf{Y}|\mathbf{W}\mathbf{H}) = \sum_{f,n} d \left(Y_{f,n} \left| \sum_c W_{f,c} H_{c,n} \right. \right) \quad (2)$$

where d is a scalar divergence. common way to measure how close \mathbf{Y} and $\mathbf{W}\mathbf{H}$ are to use a so-called Beta divergence [5], defined by

$$d_\beta(y|x) = \begin{cases} \frac{y^\beta}{\beta(\beta-1)} + \frac{x^\beta}{\beta} - \frac{yx^{\beta-1}}{\beta-1} & \beta \in \mathfrak{R} \setminus \{0,1\} \\ y(\log y - \log x) + (x-y) & \beta = 1 \\ \frac{y}{x} - \log \frac{y}{x} - 1 & \beta = 0 \end{cases} \quad (3)$$

The limiting cases $\beta=0$ and $\beta=1$ correspond to the Itakura-Saito (IS) and Kullback-Leibler (KL) divergences, respectively. Another case of note is $\beta=2$ which corresponds to the Least Square (LS) distance. The Beta divergence offers a continuum of noise statistics that interpolates between these three specific cases. This paper proposed a new model of monaural source separation based on two-dimensional NMF (NMF2D) model [6] with the Beta-divergence as an objective function. We develop a novel solution that efficiently performs source separation to be used in audio source separation. The proposed solution operated in time-frequency domain and the objective function was minimized using multiplicative update rules.

The remainder of the paper is organized as follows: Single channel mixture model in the TF domain is introduced in Section II. The derivation of proposed separation technique of Beta-divergence two dimensional NMF is detailed in Section III. Section IV presents the results of experimental tests as well as the analysis. Finally, Section V concludes the paper.

II. TWO-DIMENSIONAL NMF WITH BETA-DIVERGENCE

A. Source Model

In this section, the proposed nonnegative matrix factorization framework is derived. Firstly, we considered a source model of \mathbf{Y} which is defined as a follows:

$$\mathbf{Y} \approx \sum_{\tau=0}^{\tau_{\max}} \sum_{\phi=0}^{\phi_{\max}} \mathbf{W}^{\tau} \mathbf{H}^{\phi} \approx \sum_{\tau=0}^{\tau_{\max}} \sum_{\phi=0}^{\phi_{\max}} \left(\sum_{j=1}^J \mathbf{W}_j^{\tau} \mathbf{H}_j^{\phi} \right) \quad (4)$$

where J is the number of sources. The matrix \mathbf{W}^{τ} represents the τ^{th} slice spectral basis and \mathbf{H}^{ϕ} represents the ϕ^{th} slice of temporal code for each spectral basis element. The vertical arrow in \mathbf{W}^{τ} denotes downward shift operator which moves each element in the matrix by ϕ row down. By the same token, the horizontal arrow in \mathbf{H}^{ϕ} denotes the right shift operator which moves each element in the matrix by τ column to the right. This can be interpreted as follows, i.e:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \\ 1 & 2 & 3 \end{bmatrix}, \quad \overset{\rightarrow 1}{\mathbf{A}} = \begin{bmatrix} 0 & 1 & 2 \\ 0 & 1 & 2 \\ 0 & 1 & 2 \end{bmatrix}, \quad \overset{\downarrow 2}{\mathbf{A}} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 2 & 3 \end{bmatrix}.$$

The factorization for NMF2D source model in (4) is based on a model that represents temporal structure and pitch change. In audio processing, the model represents each instrument by a single time-frequency profile convolved in both time and frequency by a time-pitch weight matrix. This model thoroughly decreases the number components need to model various instruments and efficiently solves the monaural source separation problem. In the following, novel algorithm of sparse NMF2D with Beta-divergence is proposed to estimate the parameter of \mathbf{W}_j^{τ} and \mathbf{H}_j^{ϕ} from the mixture.

B. Cost Function with Multiplicative Update Rules.

Now, we incorporated the Beta-divergence as defined in (3) such that it will minimize the cost function as follow:

$$C_{\beta}(\mathbf{Y}|\hat{\mathbf{Y}}) = \sum_{f,n} \left(\frac{(\mathbf{Y}_{f,n})^{\beta}}{\beta(\beta-1)} + \frac{(\hat{\mathbf{Y}}_{f,n})^{\beta}}{\beta} - \frac{\mathbf{Y}_{f,n}(\hat{\mathbf{Y}}_{f,n})^{\beta-1}}{\beta-1} \right) \quad (5)$$

for $f = 1, \dots, F$, $n = 1, \dots, N$ where $\hat{\mathbf{Y}} = \sum_{j,\tau,\phi} \mathbf{W}_j^{\tau} \mathbf{H}_j^{\phi}$. In this paper, we employed the multiplicative update rules which consist in updating each parameter by multiplying its value at the previous iteration by a certain coefficient. The derivatives of (5) corresponding to \mathbf{W}^{τ} and \mathbf{H}^{ϕ} of Beta-NMF2D are given by:

$$\frac{\partial C_{\beta}}{\partial \mathbf{W}_{f',j'}^{\tau}} = \frac{\partial}{\partial \mathbf{W}_{f',j'}^{\tau}} \left(\sum_{f,n} \left(\frac{(\mathbf{Y}_{f,n})^{\beta}}{\beta(\beta-1)} + \frac{(\hat{\mathbf{Y}}_{f,n})^{\beta}}{\beta} - \frac{\mathbf{Y}_{f,n}(\hat{\mathbf{Y}}_{f,n})^{\beta-1}}{\beta-1} \right) \right) \quad (6)$$

$$= \sum_{\phi,n} \left((\hat{\mathbf{Y}}_{f'+\phi,n}^{\tau})^{\beta-1} - |\mathbf{Y}_{f'+\phi,n}^{\tau}|^2 (\hat{\mathbf{Y}}_{f'+\phi,n}^{\tau})^{\beta-2} \right) \mathbf{H}_{j',n-\tau}^{\phi}$$

and

$$\frac{\partial C_{\beta}}{\partial \mathbf{H}_{j',n'}^{\phi}} = \frac{\partial}{\partial \mathbf{H}_{j',n'}^{\phi}} \left(\sum_{f,n} \left(\frac{(\mathbf{Y}_{f,n})^{\beta}}{\beta(\beta-1)} + \frac{(\hat{\mathbf{Y}}_{f,n})^{\beta}}{\beta} - \frac{\mathbf{Y}_{f,n}(\hat{\mathbf{Y}}_{f,n})^{\beta-1}}{\beta-1} \right) \right) \quad (7)$$

$$= \sum_{\tau,f} \mathbf{W}_{f-\phi',j'}^{\tau} \left((\hat{\mathbf{Y}}_{f,n'+\tau}^{\phi})^{\beta-1} - |\mathbf{Y}_{f,n'+\tau}^{\phi}|^2 (\hat{\mathbf{Y}}_{f,n'+\tau}^{\phi})^{\beta-2} \right)$$

Thus, by applying the standard multiplicative update rule:

$$\mathbf{W}_{f',j'}^{\tau'} \leftarrow \mathbf{W}_{f',j'}^{\tau'} - \eta_W \frac{\partial C_{\beta}}{\partial \mathbf{W}_{f',j'}^{\tau'}} \quad \text{and} \quad \mathbf{H}_{j',n'}^{\phi'} \leftarrow \mathbf{H}_{j',n'}^{\phi'} - \eta_H \frac{\partial C_{\beta}}{\partial \mathbf{H}_{j',n'}^{\phi'}} \quad (8)$$

where η_W and η_H are positive learning rates which can be obtained by following [7], namely:

$$\eta_W = \frac{\mathbf{W}_{f',j'}^{\tau'}}{\sum_{\phi,n} (\hat{\mathbf{Y}}_{f'+\phi,n}^{\tau'})^{\beta-1} \mathbf{H}_{j',n-\tau'}^{\phi'}} \quad \text{and} \quad \eta_H = \frac{\mathbf{H}_{j',n'}^{\phi'}}{\sum_{\tau,f} \mathbf{W}_{f-\phi',j'}^{\tau} (\hat{\mathbf{Y}}_{f,n'+\tau}^{\phi'})^{\beta-1}} \quad (9)$$

Thus, the multiplicative update rules for \mathbf{W}^{τ} and \mathbf{H}^{ϕ} become:

$$\mathbf{H}^{\phi} \leftarrow \mathbf{H}^{\phi} \cdot \frac{\sum_{\tau} \mathbf{W}^{\tau} \left(\left(\overset{\leftarrow \tau}{\hat{\mathbf{Y}}} \right)^{(\beta-2)} \cdot \overset{\leftarrow \tau}{\mathbf{Y}} \right)}{\sum_{\tau} \mathbf{W}^{\tau} \left(\overset{\leftarrow \tau}{\hat{\mathbf{Y}}} \right)^{(\beta-1)}} \quad (10)$$

and

$$\mathbf{W}^{\tau} \leftarrow \mathbf{W}^{\tau} \cdot \frac{\sum_{\phi} \left(\left(\overset{\uparrow \phi}{\hat{\mathbf{Y}}} \right)^{(\beta-2)} \cdot \overset{\uparrow \phi}{\mathbf{Y}} \right) \overset{\rightarrow \tau}{\mathbf{H}}^{\phi}}{\sum_{\phi} \left(\overset{\uparrow \phi}{\hat{\mathbf{Y}}} \right)^{(\beta-1)} \overset{\rightarrow \tau}{\mathbf{H}}^{\phi}} \quad (11)$$

In equations (10) and (11), $\mathbf{A} \cdot \mathbf{B}$ denotes element wise multiplication and $\frac{\mathbf{A}}{\mathbf{B}}$ denotes the element wise division.

C. Reconstruction of the Separated Sources

From mixture \mathbf{Y} , we seek the two estimated sources which are $\hat{\mathbf{X}}_1 = \sum_{\tau=0}^{\tau_{\max}} \sum_{\phi=0}^{\phi_{\max}} \mathbf{W}_1^{\tau} \mathbf{H}_1^{\phi}$ and $\hat{\mathbf{X}}_2 = \sum_{\tau=0}^{\tau_{\max}} \sum_{\phi=0}^{\phi_{\max}} \mathbf{W}_2^{\tau} \mathbf{H}_2^{\phi}$. Then, by

using binary masking technique [8], we obtained mask, \mathbf{M}_j as follows:

$$\mathbf{M}_j = \begin{cases} 1, & \text{if } \hat{\mathbf{X}}_j > \hat{\mathbf{X}}_k \\ 0, & \text{Otherwise} \end{cases} \quad (12)$$

Then, the time domain estimated signal $\hat{\mathbf{x}}_j$ is obtained by resynthesizing \mathbf{M}_j with the mixture \mathbf{Y} i.e. $\hat{\mathbf{x}}_j = \text{resynthesize}(\mathbf{M}_j \cdot \mathbf{Y})$. Here, ‘resynthesize’ signifies the inverse mapping of log-frequency axis to the original frequency axis and then followed by inverse short-time Fourier transform (STFT) back to the time domain.

III. EXPERIMENTS & ANALYSIS

A. Experiment Setup

The proposed algorithm is tested on audio signals containing female speech and jazz music. The mixture is approximately 6s long and sampled at 16 kHz. For audio separation, after conducting the Monte-Carlo experiments over 50 independent realizations of the mixture, the parameters of the convolutive factors of τ and ϕ shifts are set to be $\tau_{\max} = 8$ and $\phi_{\max} = 32$. This is the best realistic parameter setting to represent the temporal code and spectral basis in the factorization for most of music signals. To evaluate this, the performances of the algorithm have been measured using signal to distortion ratio (SDR) [9] which measures an overall sound quality of the source separation. SDR value which is higher than 7dB can be considered as good because it shows that there is less distortion in the recovered signal and represents an acceptable perceptual measure.

B. Audio Source Separation Results

Figure 1 show the average SDR values obtained from various values of Beta using multiplicative update NMF2D algorithm. The value of β tested was varied from 0 to 2 in steps of 0.1. It ought to cover Least Square (LS) distant, the Kullback-Leibler (KL) divergence and the Itakura-Saito (IS) divergence of NMF2D. The average separation performance was obtained from the estimated SDR value for each source in a speech-music mixture, thereby providing a measure of overall separation for each signal. From Figure 1, as we increase the value of β , the performance also increase and it reach its peak value when $\beta=0.8$ with average SDR value of 8.5dB is obtained for each source. A tail-off in performance occurs as the value of β increases from 0.8 goes up to 2. From this experiment, it suggests that beta around 0.8 is an optimal value for audio separation. Figure 2 shows the audio separation results in time domain. From Figure 2, the proposed algorithm shows the capability to separate the single mixture and recover the female speech and jazz music very well in

single channel mixture. It can be seen that the separated signals almost replicate the original sources.

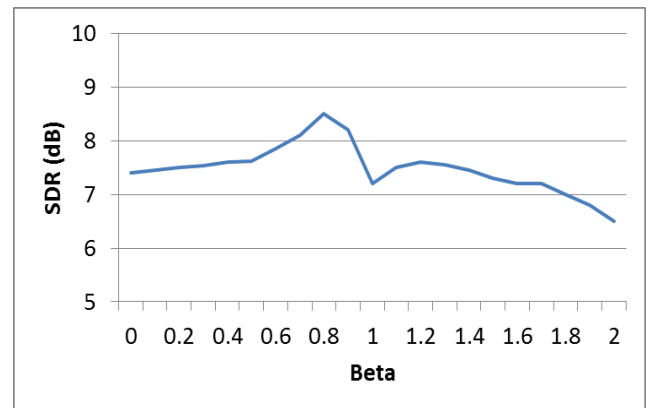


Figure 1 Separation results for various values of β using Beta-divergence NMF2D

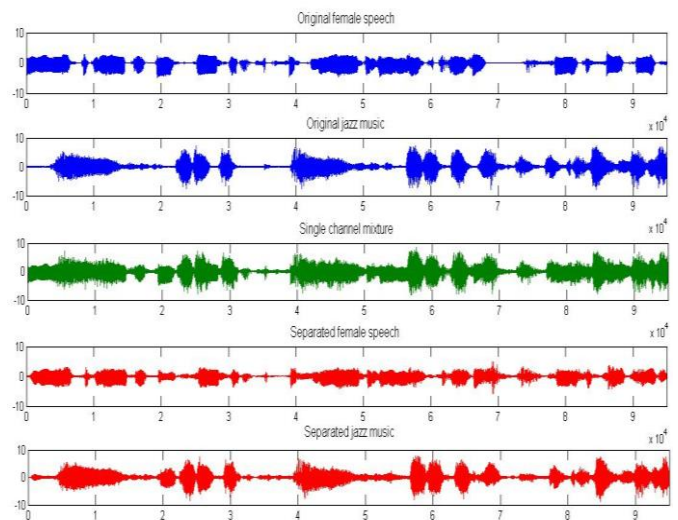


Figure 2 Audio separation results using Beta-divergence NMF2D

IV. CONCLUSION

The use of the Beta-divergence for audio source separation using NMF2D model has been investigated. The value of Beta-divergence with $\beta=0.8$ was found to produce an optimal result. The method proposed are computationally efficient where it avoids strong constrains of separating sources without prior knowledge of the original sources. We confirmed through an experiment that the proposed algorithm performs very well in separation of an audio mixture.

ACKNOWLEDGMENT

The authors would like to thank Universiti Teknikal Malaysia Melaka (UTeM) and Ministry of Education Malaysia for the

research grant funding RAGS/2013/FKEKK/TK02/04/B0033 that makes this research work possible.

REFERENCES

- [1] Li. Y, Woodruff J. and Wang D.L., 2009, "Monaural musical sound separation based on pitch and common amplitude modulation", IEEE Transaction on Audio, Speech and Language Processing, 17, 1361-1371.
- [2] Gao Bin, Woo W.L., and Dlay S.S., 2008 , "Single Channel Blind Source Separation using best characteristic basis," Proceeding of 2008 3rd International Conference of Information and Communication Technologies: From Theory to Applications, 1-5.
- [3] Darsono A.M, Gao Bin, Woo W.L, Dlay S.S, 2010, "Nonlinear single channel source separation", International Symposium On Communications Systems, Networks And Digital Signal Processing (CSNDSP), 507-511.
- [4] Kompass R., 2005 "A generalized divergence measure for non-negative matrix factorization", Neuroinformatics workshop, Torun, Poland.
- [5] Fevotte C., Bertin N., and Durrieu J.L., 2009, "Nonnegative matrix factorization with the Itakura-Saito divergence. with application to music analysis," Neural Computation, 21, 793-830.
- [6] Morup M. and Schmidt M.N., 2006 "Sparse nonnegative matrix factor 2-D deconvolution," Technical Report, Technical University of Denmark, Copenhagen, Denmark..

The Trends and Directions of Wisdom and Semantic-based Search System

Soheli Farhana, Md. Masum Billah

Faculty of Engineering
International Islamic University Malaysia
53100 Kuala Lumpur, Malaysia
soheli.farhana@live.iium.edu.my, mdmasum.b@live.iium.edu.my

Abstract— Peoples are able to share their knowledge and information through the online web. A large database is capable to handle global information through the web as well. Thus huge number of databases is grown due to store information. In this information handling cases, it is needed to search by using dedicated tools; broadly known as search engine. Though a number of search engines are on hand these days but recovering of authenticated information is quite complicated now. Furthermore, these existing search engines are not able to indicate the authenticated and doubtful retrieved information. On the other hand, to overcome these limitations in investigating systems to rescue authenticate data smartly; wisdom and semantic web search systems are performing main responsibility. An intensive literature survey on the search engine in wisdom and semantic search technologies are presented in this paper.

Keywords— *data rescue; wisdom and semantic search, search systems.*

I. INTRODUCTION

"Wisdom" is a word which contains the meaning of the thoughtful theory and function of awareness. It is a full of meaning accepting and awareness of persons, possessions, dealings or situations, consequential in the ability to apply perceptions, judgments and actions in keeping with this understanding. It often requires control of one's emotional reactions (the "passions") so that universal principles, reason and knowledge prevail to determine one's actions. Wisdom is also the comprehension of what is true coupled with optimum judgment as to action (wikipedia). The Wisdom web is a such kind of search system which will perform the authenticate data retrieval system from the authorized database. However, wisdom-based search systems are not revealed vastly in the current search systems.

Additional room of current web is Semantic Web which permits the precise data explained with such terminologies that can be easily understand by human and intelligent machines [1]. An updated W3C model is known as Resource Description Framework (RDF) is used in description of semantic web information. Human and computer could be collect available data from semantic web by using current web sites. Semantic web contains significant concept named ontology [2]. Web Ontology Languages (OWL) is another W3C model is used for ontology representation. Existing web systems cannot resolve the internal operational problem where semantic web can able to demonstrate efficiently data detection, computerization and incorporation. Semantic search systems are still in primary phase considering in research scale, whether the existing web search systems such as msn, yahoo, google etc are still control the web world. Maximum search

systems are used keyword to the user input data search in the web page to retrieve. Nevertheless they use sophisticated algorithms to verify the real information searching from the pointless webs. It can able to reply according to the subject of the searching information. On the other hand, lack of their web information dependency, it fails to reply answer of the intellectual enquiries from client. Presenting a semi-precise in less delay is the key goal of this kind of search systems. Therefore most of the users are not satisfied using such kind of search systems. In addition, such systems are enable to retrieve and verify the information from a authenticated source. Thus, semantic web systems are able to be dealt with intelligent enquiries [3] and wisdom-based system be able to dealt with authenticated sources.

Trends and directions of the wisdom and semantic-based search systems are discussed and analyzed in this work.

II. STATE OF ART

Data search and rescue is from the web contains a lot of challenges for the basic data retrieval. For the different types of technologies using in different web search systems, the retrieved results may be vary with their expectation. Though this kind of search systems only retrieve the data from the web only, but currently few researcher are interested to develop semantic-based search systems. Existing web database got lack of semantics' systems that turn it in more difficulties to understand the client expectations. Three types of problem arise during the data surrounding on the web. First, how the search systems make a plan to search data on the web in a intelligent and useful manner, which is important to the user. Second, how the search systems can distinguish the search

results to the several web link. Graph-based query model can solve the second problem [4]. Third, do the search results authenticate? By producing semantic web annotation, it can solve the first stated problem [5-6]. Authenticated ontology research is still in early stage to solve the third problem. Semantic web layer working principle is shown in Fig. 1.

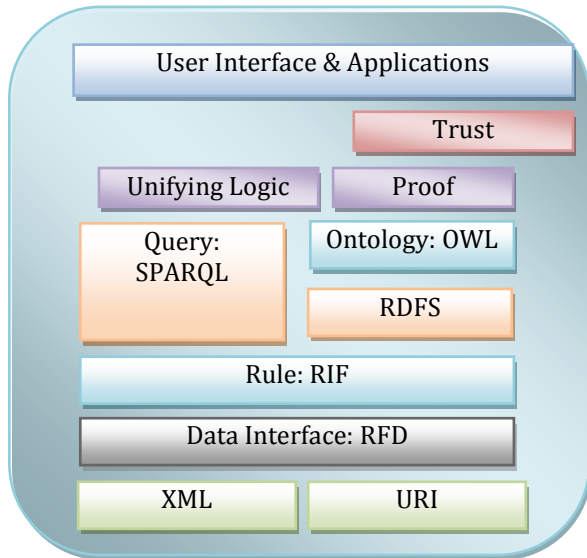


Fig. 1. Structure of Semantic Web

A. Existing Web Limitations

Lack of uses of semantic technology in existing World Wide Web causes search systems ambiguous search result. Therefore semantic and wisdom-based search systems can be defeat some problems; such as, lack of proper retrieved data representation structure, ambiguous problem in proper search results representation, lack of intelligent data distribution, etc.

One of the best semantic search engines is 'Hakia' can be able to perform structured search [7-8]. The strategy of this search is to depend on the meaning of the queries. All the features of these search systems are used of semantic technology, which can be produced digital object [7], [9].

III. WISDOM BASED SEARCH SYSTEMS

A. Search Systems

Few numbers of wisdom-based search systems are designed for several working situations and the techniques that understand are diverse. A semantic base search system is designed for perform automation search systems by combining digital ontology and description logic inference system [10]. This search system presents a recipe which is able to formulating the demands of wisdom search system and makes a solution of the efficiency of search system. To incorporate

ontology library with the client assumption by using Descriptive Logic Inference System [11] which enable the search system to accomplish the complete search for wisdom-base search system.

Only texts are used for searching purpose on a web by the maximum existing search systems. Some operation could be done by the representative for a client of the computer. Every search systems' representative helps to assist their every client of the system. To propose its own client and communicate with other representatives is the main objective for the representative of a client. This representative may be used for various external sources information. Actually this representative [12] is software made representative which is working in the server system. A precision augmentation search engine was developed for retrieval of information which tries to follow the techniques of wisdom-based search systems [13]. The default and assumption of the information was used in this technique. The default information was used for search information's' returns data that comprise distinctive comfortable data of a issue. Information assumption was used for search information's' returns data that comprise data close to expected enquiries. They have implemented this system using a limited database with a potential returns. The analyses were conducted using fuzzy and heuristic satisfaction function [14]. Another intellectual methods was used for authenticate search system depends on client inclination [15]. These systems are very helpful to different client data search system for quality assurance of the retrieve data. Another researchers are presumptuous a search system that perform general mechanism for wisdom-based search system. Results are verified in a practical method of following the supply managing of data consistent with intangible systems [16].

IV. SEMANTIC SEARCH SYSTEMS

Huge meaning of information from a search result using a special process is called semantic system. Semantic systems contain a series of code that are used to converse denotation, and this converse could be influence the behaviour. Future generation web has been driven by semantic web. An indication has been made of 'Semantic' that the significance of information on the web can be discovered not just by individuals, but also by computers. Then the Semantic Web was created to extend the web and make data easy to reuse everywhere. Semantic web is being developed to overcome the following main limitations of the existing Web [17]; such as, lack of proper data representation problem, weak data intercommunication represent the ambiguous information, lack of data transformation automation, not able to communicate huge number of user and not capable to capture the global format of the data.

At present numerous semantic search systems are executed in diverse running situation. This system may be keep using to comprehend the existing search systems. Semantic systems keep data of websites possessions which enable it to resolve compound enquiries, taking into consideration to the circumstance where websites possessions are beleaguered [18].

A huge contribution is done by semantic systems over the improvement of digital applications [19]. General query language is used in this system which proceeds semantically associated data remains to assure the clients' enquiry using XSearch system. The presentation of the diverse methods and also the reminder and the accuracy had calculated experimentally. XSearch efficiency is justified in terms of scalability by this experiment. Another proposed Semantic-based search systems was proposed to use an assumption replica to construct the contacts among records [20]. This system has two divisions: first is the crawler division. Here files are extracted from file system to produce two indices: the file rank metrics will ranked in index and the other one is the keyword index for recording the keywords. The second division is evolved with the query system using the keyword corresponding with the search item to identify the similar files. Then File Rank is used to rank the query files. A semantic search system was developed by Wang et al. to extract information from tables with the following steps: mark the table cells semantic relation, convert and store the table information into the database and using query language extract the objective data [21]. Avatar was developed by Kandogan et al. for the text search engine using the ontology [22]. Avatar consists of two functions, UIMA frame work and automatic transforming of the interpreting keywords. Fig. 2 shows the working flow of the AVATAR search engine.

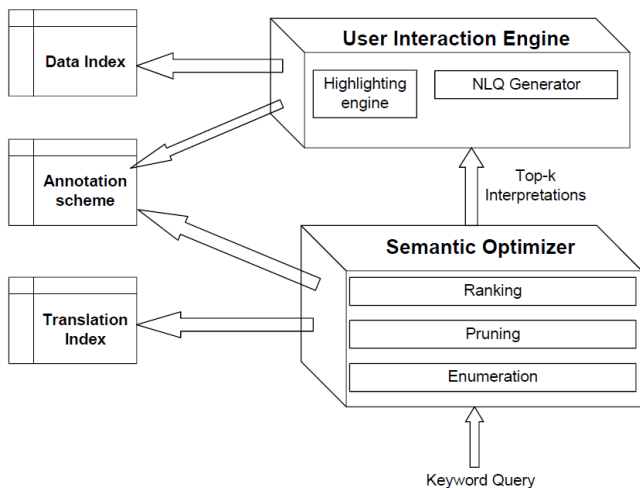


Fig. 2. Working flow of AVATAR search engine [22]

V. ONTOLOGY BASED SEARCH SYSTEMS

Ontology search system was developed by Maedche et al. [23]. An ontology registry is premeditated to accumulate the data about ontology in its architecture. In ontology registry, ontology search is operated in two conditions query by example is to query by term is to restrict the hyponyms of terms for search and restrict search fields. The search system is shown in Fig. 3 for stating the whole process of the search system.

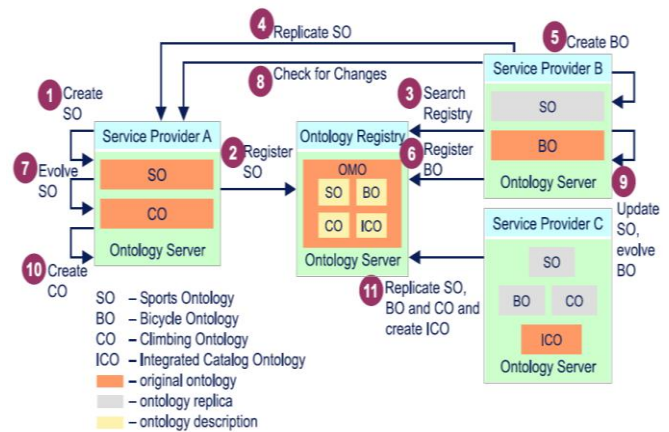


Fig. 3. Working flow of Ontology Search Engine [23]

VI. GENERAL CONCERNS

A review is conducted on the current and vibrant part in wisdom and semantic search systems. Some general issues are tinted in the existing semantic and wisdom search systems are done as follows:

A. Less accuracy and maximum reminder

Few of the semantic search systems are not able to demonstrate their important presentation for upgrading the accuracy. An experiment was done in a search system [24-25] which present the less accuracy and maximum reminder during returning the search result.

B. Client Identification Intention

Semantic search system contains client identification. A method was developed for investigating the demand provisions set the client identification intention, thus the search system activity will be more appropriate for the client.

C. Ambiguous returns

Client can be key in an ambiguous word which will result the wrong reply from the search system. For example, client type java but the system returns coffee bean etc.

D. Incorrect enquiries

Clients are sometimes plays are wrong role with the search system by key in the improper keywords. Thus the system will return irrelevant results.

E. Authenticated Information

The retrieved data from semantic or wisdom-based search systems are not claimed 100% precise correct data which are authenticated by real source.

VII. CONCLUSIONS

A brief survey of the existing literature regarding wisdom and semantic web search system is discussed in this paper. A short review is done against those features correspondingly. Additionally, the concerns in the surveyed semantic and wisdom search systems are over and done with five viewpoints within the programmer and clients' awareness, techniques, less accuracy and maximum reminder, short of experiment and information authentication.

REFERENCES

- [1] A. Maedche and S. Staab. "Ontology learning for the semantic web." *Intelligent Systems, IEEE* 16.2 (2001): 72-79.
- [2] K. Wolstencroft, "RightField: Scientific Knowledge Acquisition by Stealth through Ontology-Enabled Spreadsheets." *Knowledge Engineering and Knowledge Management* (2012): 438-441.
- [3] J. Steinberger, and K. Ježek. "Evaluation Measures for Text Summarization." *Computing and Informatics* 28.2 (2012): 251-275.
- [4] H. Hao, "BLINKS: ranked keyword searches on graphs." *Proceedings of the 2007 ACM SIGMOD international conference on Management of data. ACM*, 2007.
- [5] M. Fox, and J. Huang. "Knowledge provenance: An approach to modeling and maintaining the evolution and validity of knowledge." *University of Toronto* (2003).
- [6] F. Almeida and J. Lourenço. "Creation of value with Web 3.0 technologies." *Information Systems and Technologies (CISTI), 2011 6th Iberian Conference on. IEEE*, 2011.
- [7] S. Faizan, et al. "SWISE: Semantic Web based intelligent search engine." *Information and Emerging Technologies (ICIET), 2010 International Conference on. IEEE*, 2010.
- [8] T. Valentin, D. Damljanovic, and K. Bontcheva. "A natural language query interface to structured information." *The Semantic Web: Research and Applications* (2008): 361-375.
- [9] H. Junguk, et al. "Ontology-based Brucella vaccine literature indexing and systematic analysis of gene-vaccine association network." *BMC immunology* 12.1 (2011): 49.
- [10] S. Sheng, et al. "A framework of service oriented semantic search engine." *Computational Problem-Solving (ICCP), 2011 International Conference on. IEEE*, 2011.
- [11] S. Inamdar and G. N. Shinde "An Agent Based Intelligent Search Engine System for Web mining" *Research, Reflections and Innovations in Integrating ICT in education*. 2008.
- [12] L. Zhan, L. Zhijing, "Web Mining Based On Multi-Agents", *Computer Society,IEEE*(2003).
- [13] R. S. Raja, M. Marrikkannan, and S. Karthik. "Need Of Future Web Technology The Semantic Web A Brief Survey." *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)* 1.10 (2012): pp-71.
- [14] S. P. K. Satya and S. V. Raghavan "Intelligent Search Engine: Simulation to Implementation", In the proceedings of 6th International conference on Information Integration and Web-based Applications and Services (iiWAS2004), pp. 203-212, September 27 - 29, 2004, Jakarta, Indonesia, ISBN 3-85403-183-01.
- [15] D. Meng, X. Huang "An Interactive Intelligent Search Engine Model Research Based on User Information Preference", *9th International Conference on Computer Science and Informatics, 2006 Proceedings, ISBN 978-90-78677-01-7*.
- [16] X.S. Yan, X. Junyang, Y. Zhang "Intelligent Search Engine Based on Formal Concept Analysis" *IEEE International Conference on Granular Computing*, pp 669, 2-4 Nov, 2007.
- [17] S. kumar, S. k. malik "Towards Semantic Web Based Search Engines" *National Conference on "Advances in Computer Networks & Information Technology (NCACNIT- 2009) March 24-25*,
- [18] F. F.Ramos, H. Unger, V. Larios (Eds.): *LNCS 3061*, pp. 145-157, Springer-Verlag Berlin Heidelberg 2004.
- [19] C. S. Mamou, J. Kanza, Y. Sagiv, Y "XSEarch: A Semantic Search Engine for XML" *proceedings of the international conference on very large databases*, pages 45-56, 2003.
- [20] D. Bhagwat and N. Polyzotis, "Searching a file system using inferred semantic links," in *Proceedings of HYPERTEXT '05 Salzburg*, 2005, pp. 85-87.
- [21] H. Dong, K.H. Farookh and C. Elizabeth, "A service search engine for the industrial digital ecosystems." *Industrial Electronics, IEEE Transactions on* 58.6 (2011): 2183-2196.
- [22] E. Kandogan, R. Krishnamurthy, S. Raghavan, S. Vaithyanathan, and H. Zhu, "Avatar semantic search: a database approach to information retrieval," in *Proceedings of SIGMOD '06 Chicago*, 2006, pp. 790-792.
- [23] A. Maedche, B. Motik, L. Stojanovic, R. Studer, and R. Volz, "An infrastructure for searching, reusing and evolving distributed ontologies," in *Proceedings of WWW '03 Budapest*, 2003, pp. 439-448.
- [24] D. Ding, J. Yang, Q. Li, L. Wang, and W. Liu, "Towards a flash search engine based on expressive semantics," in *Proceedings of WWW Alt.'04 New York*, 2004, pp. 472-473.
- [25] C. Hon, L. Alan, "Toward Intention Aware Semantic Web Service Systems," *scc*, vol. 1, pp.69-76, 2005

Evaluating the Success of Information Strategic System Planning

(Two Cases from Jordan)

Qais Hammouri

MIS Department, IT College
Yarmouk University
Irbid, Jordan
hammouriqais@yahoo.com

Laith Shraideh

MIS Department, IT College
Yarmouk University
Irbid, Jordan
Laith.shraideh@hotmail.com

Emad Abu-Shanab

MIS Department, IT College
Yarmouk University
Irbid, Jordan
abushanab@yu.edu.jo

Abstract— Strategic Information System Planning (SISP) is a very important issue for contemporary organizations, where information technology is becoming an important driver of sustainable competitive advantage. This paper focuses on the success of the planning process by exploring its proceedings in two cases in Jordan. The paper utilized Earl's Model and applied it on two cases in Jordan by exploring and analyzing the strategic planning process for the two firms "Japan Tobacco International" and "Irbid Electricity Company". Results indicated that evaluating the success of SISP in JTI is more effective and focused than IDECO based on a combination of factors. The clarity of strategy, the stakeholders influence and nature of "social behavior", and the competitive environment are three main factors that support the process of evaluating the success of SISP in the organization.

Keywords—(SISP); Planning; Alignment; Public Sector; Private Sector

I. INTRODUCTION

Strategic information systems (SIS) are important applications that support the competitiveness of contemporary organizations. They are (SIS) a vital condition in business environment for gaining a competitive advantage as they are an integral part of the strategic information system planning structure for the purpose of achieving success [12].

Strategic Information System Planning (SISP) is a critical management issue as it plays an important role in helping organizations explore the environment, monitor the new development in IT, watch for competitors' actions (related to IT in markets), and improve business objectives and functions to achieve business needs utilizing IT [24]. SISP is the main

process in the alignment of information systems and business needs [32], where this strategic alignment plays a significant role in the success of SISP [5]. Without the effective IT/business alignment, businesses will not gain competitive advantage in their markets [13].

The opportunities open for success are improved by using various information technology applications with the appropriate planning process. SISP is an important approach that helps organizations utilizes their resources effectively to achieve their business goals and gain a competitive advantage.

Top management should be aware of the appropriate methodologies to ensure the success of SISP in their organizations [4].

This paper will explore the practices utilized by two Jordanian firms in their SISP process. The cases explored are Japan Tobacco International (JTI), and Irbid Electricity Company (IDECO). The structure of paper is as follows: a literature review was conducted to understand the SISP concept and the strategic environment. The following two sections covered the details of SISP approaches within the two cases. Finally, each case included a conclusion section related to its details.

II. LITERATURE REVIEW

Reference [26, p: 446] defined SISP as “*the process of identifying a portfolio of computer based-applications that will assist an organization in executing its business plans and consequently realizing its business goals*”. According to reference [7, p: 17], SISP is also “*the process of strategic thinking that identifies the most desirable IS on which the firm can implement and enforce its long-term IT activities and policies*”.

The integration of IS infrastructure represents a major problem in complex organizations because ISs are characterized as heterogeneous and fragmented [25]. By establishing SISP in complex organizations, it helps integrate infrastructure components effectively because SISP has a positive relationship with technical and data integration, and diverse functionality of applications [9]. IT is important to understand that utilizing IT diverse application is related to strategic planning and strategic context, where specific hardware and software are suitable for certain situations [2].

A. The Success of SISP

The success of SISP in organizations is not determined by a specific method or procedures or tool, but comes from the awareness of top management of how to manage their resources and understand the nature of their organization [4]. SISP should include a method, and a process of implementation that are considered as necessary conditions to ensure success [10]. The success of SISP in organizations is based on the process that is used in developing SISP, the strategy which has a specificity and comprehensiveness in its framework, and is a strategy that can contribute in providing a positive impact on the success of SISP [32].

Reference [10] identified a set of approaches to develop SISP in organizations and as follows: business led approach, method driven approach, administrative approach, technological approach, and organizational approach. Reference [23] asserted that the choice of the proper methodology for developing SISP provides a positive indication of success of SISP in organizations. According to a study among United Kingdom organizations, the combination of SISP development approaches increased the level of success of SISP [4, 31, 33]. Research indicated that using more than one methodology in SISP implementation provides a significant contribution to the success of SISP.

Reference [32] identified a set of situational factors that influence SISP environment and they are: market hostility, market dynamism, organizational formalism, centralization, role of IS, SISP goal, maturity of planning processes, IS participation in business planning, frequency\consistency, acceptance of plans. Also, they identified the SISP process configuration variables and they are: senior management involvement, resources, team involvement, participation, SISP imitator, influencer, IS role, method, SISP planning horizon, SISP scope, environmental assessment, comprehensiveness, flow, design focus, implementation. The authors concluded that both sets of variables are closely related and would significantly influence the success of SISP. Results of their research indicated that market hostility has a positive impact on four SISP process configuration variables and they are: resources, participation, influencer, and comprehensiveness. SISP goals have also positive impact on four SISP process configuration variables: SISP planning horizon, environmental assessment, comprehensiveness, and flow. Based on that, market hostility and SISP goals have the most positive impact for success the SISP based on SISP process configuration variables. Finally, and according to reference [29] work, the comprehensiveness of information system panning phases (strategic awareness, situation analysis, strategy conception, strategy formulation, strategy implementation planning) has a significant influence on the success of SISP.

Organizational learning is an important aspect in strategic planning, where it was evidenced in companies that relied on ad-hoc basis in planning activates. Such result was associated with the existence of non-formal method within an organizational approach [10]. Organizational learning is a key success factor in SISP when concentrating on knowledge, skills and attitude for any member of the SISP team. Also, organizational learning has a positive impact on organizations which have a past experience and a mature IT/IS stage. By increasing their SISP performance, facilitating information exchange, and improving the capabilities of their members, firms can enhance the SISP practice through a consistent decision-making process to gain competitive advantage [1].

B. The Failure of SISP

Research reported some of the unsuccessful features of SISP like: inadequate implementation, lack of top management support, time constrains, poor user-IS relationship, and resource constrains. SISP process is characterized differently based on different types of resources within an organization [10]. Such reasons are common in all causes of failure of IT projects [3][35]. Resources are classified into IT resources, human resources, financial resources, and others. Reference [24] proclaimed that financial resources are the most critical and harder for managers to determine accurately and most projects fail due to financial resources constrains. Other reasons for failure are reported by research like: lack of integration in a global organization may affect its business strategy because of the differentiation in culture, decision style, communication style, and leadership style [27]. Using

one methodology for implementing SISP might fail the SISP process [4].

According to a study in Australia related to the main reasons leading to failure of SISP, the lack of commitment by senior manager represents the main reason for failure of SISP formulation and implementation. The study also concluded that budget limitations is the main reason causing failure of SISP formulation, and the lack of alignment with business objectives is the main reason for failure of SISP implementation [30]. Research also indicated that the excessive commitment by top management and the excessive involvement by senior management and team members may have a negative impact, because they cause a waste in time and resources and take more time exploring all details to take decisions; such process leads to failure of SISP development [6].

Finally, in the Jordanian environment, research concluded that the top four reasons for IT project failure in a descending order were: poor planning, unclear objectives, and changing objectives during execution, and lack of executive support [3]. The study utilized responses from 95 IT specialists from diverse companies in Jordan.

C. SISP dimensions and alignment

There is a gap between SISP successes and IS capabilities, some researchers provided some dimensions to measure the success of SISP practices. The first dimension is the alignment scale which associate IS and business function. The second dimension is the analysis scale, which explores all related activities in the organization. The third dimension is the cooperation scale, which integrates the business functions together. The fourth dimension is improving the capabilities of learning. The last dimension is the contribution scale, which covers the improvement of different objectives (decision making, effectiveness, efficiency...) in an organization [24]. Reference [10] concluded that the alignment between information system and business needs is the most beneficial aspect of SISP. Reference [13] proposed an integration model to ensure the alignment of IS/IT with business needs. The integration model provides a set of benefits for businesses such as evaluating what the organization needs by concentrating on resource based theory, enhancing the alignment process through planning the activities, and improving organizational learning.

III. RESEARCH METHOD

This study followed a case study approach, where two cases were selected to apply the Earl (1990) approach for SISP. The two cases were utilized to understand better the evaluation and success of strategic information system planning (SISP) by using Earl Model, to determine the differences of the two cases in their SISP approach. The following two sections will explore the two cases in details and their SISP practices. Qualitative analyses on responses were applied to better understand the planning process and conclude to the research objectives. The main objective of this research is to

understand the SISP process followed by Jordanian firms. The study utilized two major cases: Japan Tobacco International (JTI), and Irbid Electricity Company (IDECO).

IV. JAPAN TOBACCO INTERNATIONAL

Japan Tobacco International (JTI) founded as a partnership of Japan Tobacco and RJ Reynolds, where they form a group of private companies operating in 120 countries in the world and Jordan is one of them [18]. The goal of JTI is to be the most successful and respected tobacco company in the world. JTI has a corporate strategy to increase profit through establishing outstanding brands, enhancing productivity and focus on continuous improvement.

The data collected in this study was collected from an online survey (questionnaire) sent to JTI-Jordan. The questions were answered by the manager of JTI in Jordan (based in Amman, Jordan). The research questions addressed in this study are adopted from Earl's work (1990). The following sections will depict the qualitative data collected and conclusions of this research.

A. IT in JTI

JTI follows a strategy where they focus on how ends (goals) will be achieved by means (resources). Aligning IS with business needs is important for SISP [10], and this objective comes first in ranking followed by seeking for competitive advantage. JTI sees technology as an important tool, where they described two systems that are implemented: The first one is a track and trace system which prevents illicit trade for supply chain, and provides support to develop and implement the anti illicit trade (AIT) technology to prevent the illegal (bootlegger) sale of its tobacco products. This system is also used for investigating where a realistic product may be delivered from within the legitimate supply chain to an unintended market [19].

The second system is the product authentication system which provides an authentication for its products. It is also a digital tax verification system that allows customers to check if the package is realistic or not by an SMS or telephone call. JTI employed tagging on their products (especially chemical products) by using a reader to vitrify them. JTI also adopted security programs to monitor their products from theft during transportation stage and monitoring the finished goods at factories and warehouses. Finally, in November 2014, JTI completed an acquisition of e-cigarettes brand, E-Lites which was defined as “consumer products that provide an inhalable vapor by direct electrical heating of a liquid contained within the device or a replaceable cartridge.” Such step was done to improve the performance of consolidated group and their cash flow [22].

B. The benefits and problems of SISP in JTI

Responses collected indicated that the main benefit of SISP is the documentation and communication describing the organization's strategy and how it should be implemented.

Responses showed that the strategic planning is analytical in nature, where managers should act strategically and concentrate on intuition and creativity [28]. JTI strategy formation itself involves synthesizing via strategic thinking, where it is considered a critical success factor and one of the key elements for JTI. As such, strategic planning occurs around the strategy formation activity.

Strategic planning has been criticized for attempting to systemize strategic thinking and strategy formation, because strategic thinking needs to synthesize (initiation and creativity) and strategic formation act as an instrument to evaluate the quality of strategic thinking in an organization [11].

C. The content heading of JTI strategy

The IS strategy of JTI included (Figure1) four stages: First is the process, where it includes four sub stages:

1) *The overview stage*, provides an overview about the types of brands for each department along the 120 countries, this stage also provides characteristics and quality for the product to be delivered to customers and to meet their expectations.

2) *Inputs*, in this stage the tobacco is imported like other materials such as papers.

3) *Activities*, to generate outputs.

4) *Output*, to produce the finished good.

The second stage is tools and approaches; JTI tries to make raw material available as input resources. JTI signed contracts with farmers to provide the best quality raw material. The third stage is strategic planning vs. financial planning; this phase is related to financial strategic plan, where financial planning is an important part of strategic planning. [24] Asserted that the financial resources are the most critical and hardest for managers to determine accurately and most projects fail due to financial resource constrains. The fourth stage is strategic planning vs. strategic thinking; this stage is considered as the key element of IS strategy for JTI and held more criticism because it includes a synthesis and intuition process.



Fig.1: The JTI IS strategic process

D. Developing IS strategy in JTI

Many of methods are used in developing IS strategy. JTI used two main methods: firstly, the balance scorecard method to align business activities to its corporate strategy to achieve its goals. The second is the strategic map method for monitoring their strategy. The two methods are used according to their business needs. Other methods are used sometimes to develop their strategy; the following are a short list [20]:

- ✓ Stages of growth used to develop its marketing strategy,
- ✓ Business system planning (BSP) to recognize business mission and objectives, and functions to determine business process for its needs.
- ✓ In-house IS strategy method and In-house business strategy method are utilized by understanding all business objectives from all separated functions and departments by the center of excellence (COE) in JTI. Also, the global development center (GDC) is responsible for identifying the direction of its strategy.
- ✓ In-house application search technique method is used by implementing enterprise resource planning system (ERP) to integrate new acquisitions, such as human resource operations, and by implementing track and trace system, all of these methods are used according to business needs.

SISP in JTI is connected to other business planning processes by coordinating the planning efforts and measuring the progress on strategic objectives and goals. Strategies should be reviewed from time to time to evaluate them, develop and improve to sustain competitive advantage. JTI reviews its IS strategy by benchmarking it to competitor or by comparing prices in the market, and may be used the same way for evaluating if SISP succeeded or not.

E. Successful SISP in JTI

JTI utilized few techniques and supporting tools for their methodology and they are: PEST analysis, Porter's five forces model, the growth share matrix, scenario planning, and SWOT analysis. PEST analysis (political, economic, social and technical) is an example of some practices conducted like: eliminating illicit trade by making an agreement with national government entities (political direction). An example is the agreements with the European Commission and EU Member States (December 2007). From the economical aspect, JTI monitors the international economic growth for the purpose of future investments. The social aspect is represented by the focus on reducing child labor in the world and contributing to establishing an international wide policy. Finally, the technical aspect is demonstrated by implementing the track and trace system and security systems for monitoring its supply chain from illicit trade [19].

JTI uses Porter five forces analysis to understand threats of new entrant by monitoring the markets, threat of substitute products or services by continuously improving their goods, the bargaining power of customers by providing loyalty to

their products, the bargaining power of suppliers by supporting the farmers of raw materials, and finally the intensity of competitive rivalry by sustaining a competitive advantage through innovation [34]. By keeping the channels of communication and discussion open for new ways and ideas, and creating opportunities for meetings to share knowledge and new ideas across the organization, JTI emphasizes knowledge and learning as a fertile ground for innovation [21].

JTI uses also the growth share matrix to evaluate its products. This matrix divides products into four categories: the stars, the cash cow, the dogs, and the question mark [8]. The first category is stars; Rolling Tobacco is one of the top rolling brands and a leader in the world and Hamlet Cigar is the leader in Greece. The second category is cash cow; Winston product has the best market share and cash revenue, where it is sold in over 100 countries. The third category is the dog; JTI offer Snus products for non smoking people, where this product has a low market share and sells only in Sweden with a low growth rate. Finally, the question mark such as the Mevius product; it's a new product and needs some time to increase its market share [22].

JTI uses scenario planning or scenario thinking to make predictions about future events and depict how the future might look like. JTI also uses SWOT analysis method to evaluate the strengths, weakness, opportunities, and threats in its environment and industry. Strengths of JTI in acquiring the leading e-cigarette brand E-Lites, the weakness of JTI is in the lack of integration between departments because of the differences in culture, decision style, communication style and leadership style [27]. Also, as government regulations try to ban JTI brands, JTI seeks to increase growth in market share as an opportunity open for it, the core revenue grew 3.3% during July-September and 2.4% during January-September in 2013. Finally, the threat to JTI is in illicit trade of its supply chain processes and industry contraction [17].

F. Conclusions

We can conclude that JTI succeeded in implementing SISP in their strategy based on Earl Model. The strategy of JTI is completed through four stages (process, tools and approaches, strategic planning vs. financial planning, and strategic planning vs. strategic thinking). IT is a critical part of the firm's strategy, and SISP is very important to achieve business needs by aligning IT with business strategy of a firm. SISP is a mechanism to become a leader in market by gaining a competitive advantage and increasing market share.

Strategic planning has been criticized for attempting to systemize strategic thinking and strategy formation, are inherently creative activities involving synthesis, and strategic planning vs. strategic thinking is the most key elements in the firm strategy. To ensure the success of SISP, the firm is implementing additional set of tools and techniques such as PEST analysis, porter five forces analysis, growth-share

matrix, scenario planning, SWOT analysis, and balance scorecards [4].

Introducing the e-cigarette brand to their products is a critical success factor which contributed to the improvement of the company's strategy to increase the competitive strength against its competitors. Based on that, we can assert that government's intervention is one of the problems that hinder company's strategy.

V. IRBID ELECTRICITY COMPANY

Irbid Electricity Company (IDECO) was founded in 1957. The vision of company is to become leader in providing electric service with high quality and distinctive specifications that are compatible with the best international standards by 2015. The mission of IDECO is to contribute to the continuity of economic and social development through providing excellent service with high quality according to international standards in all parts of company's business areas. This is done through the commitment to invest in the development of human element in order to raise efficiency and the development of capabilities to achieve better service and return that meet customer needs and exceed the expectations of employees and all stakeholders involved in the company [14].

According to a Forbes study that included 324 companies from Jordan, Saudi Arabia, Egypt, Kuwait, UAE, Bahrain, Oman, Qatar, and Lebanon, results indicated that the administrative stability in IDECO contributed to the increased profit and enhanced financial performance during the past five years. The study showed that the IDECO had the strongest executive managements based on four standards: return on shareholders in 2011, the rate of market share, the rate of asset growth, and the rate of earnings growth per share. The Forbes study stated that the general manager of IDECO is the most powerful general manager in Jordan within the energy field [15].

The data collected for this case through an interview with the manager of IT department in the company. The interview was built around a survey (questionnaire) that took one hour of time. The items included in the interview are adapted from Earl (1990), and similar to the set used for the JTI case. Results of the survey indicated that IDECO strategy is derived from the government and there is no clear strategy because the process and activities are setup and implemented based on day to day needs. The commitment of top management is very important for IDECO and respondent declared that top management commitment is the main objective for developing IS/IT strategy. Also, IT department provides a technical support rather than aligning IT in the strategy of a firm, where the respondent described the IT department involvement within the strategy of IDECO as part of a workflow system.

A. IT in IDECO

Workflow system is the first completed ERP system in Jordan that is developed in house to follow up all activities and orders

in a company. The workflow system provides different set of activates such as, manage the in/outgoing transactions, executing reports, managing and executing all types of transactions such as governmental transactions and in house transactions. This system also identifies a profile for each employee in the company under the name of employees' services that contain all the information needed related to employees and provides different services such as request of vacations.

Workflow system also is used for planning the resources of organizations and help to query about the archived files (electronic archiving). Also, it helps query about the amount of materials in warehouses and provide an overview about emergency breakdowns in any location. Workflow system has a renewable system; it's a transaction system that helps to provide queries about the bills of services, and information about the subscriptions of customers. The CIO declared that the workflow system is considered as an application that presents an opportunity to sustain a competitive advantage for IDECO.

IDECO is the first electricity distribution company in Jordan that activate the electronic bill payment "E-payment" in collaboration with the Central Bank through a set of tools and techniques such as net-banking, mobile banking, and ATM. Such step aims to facilitate the payment of bills from any place in Jordan and at any time without charging extra commission [16].

B. Developing IS strategy in IDECO

The responses collected in relation to the SISP process in IDECO declare that the methods used for developing there is strategy are the following: the first method is the stages of growth model, where it develops the workflow system in step by step process according to their business needs. The second method is business system planning for monitoring business channels. The third method is enterprise modeling for new and change management. Finally, the fourth method is information engineering which integrates data with other departments of IDECO. IDECO also used In-house IS strategy, In-house business strategy, and In-house application search technique for the development of the workflow system according to their business needs.

Due to the nature of company's work and its direct relation to citizens, the social behavior and the nature of people are the main problems that hinder the success of the company's business strategy. The company faces some difficulty in dealing with some people and convincing them to pay the bills of the services provided by IDECO.

IDECO is the only company that offers the electric services in Irbid city, this leads to the absence of a competitive environment, which have a key role for continuous improvement of company's strategy.

C. Conclusions

Compared with IDECO, JTI succeed in implementing SISP in their strategy based on Earl Model, by analyzing its strategy, and evaluating its strategy methods and tools that are considered as a cornerstone to the success of SISP. Although IT is a major driver in its routine processes, IDECO could not identify the success of SISP because of the lack of clarity in its strategy, where its strategy is derived from the government directions, in the other hand, the absence of a competitive environment, and a negative impact of social behavior. Based on such direction, IDECO failed to implement SISP based on Earl Model, which might not be a downside rather than an opportunity to future research to map other schemes and approaches to SISP. The public perspective is dominant in the case of IDECO, where the Jordanian government perspective prevails as a major stockholder in the firm.

Evaluating the success of SISP in both private and public sectors needs to take many cases into consideration which consumes more time, but this study was limited to two cases only. Through these case studies, it is visible that evaluating the success of SISP in private sector is an easier and more focus process than in public sector. Such conclusion is based on three factors: the clarity of strategy, the stakeholders influence and nature, and the competitive environment

REFERENCES

- [1] F. Abu Baker, M.A Suhaimi, and H. Hussin, "Evaluating strategic information system planning (SISP) among Malaysian Government Agencies using organizational learning-based model", *The European Conference on Information Management and Evaluation*, 2013, pp. 45-53.
- [2] E. Abu-Shanab, "Data warehousing strategies: a strategic alignment perspective". A conceptual paper presented in the *Proceedings of the Decision Sciences Institute Conference-Boston*, USA, 2004, pp. 1-6.
- [3] E. Abu-Shanab and A. Al-Saggar, "Reasons behind it project failure: the case of Jordan". A book chapter in "*Business strategies and approaches for effective engineering management*" edited by Saqib Saeed, M. Ayoub Khan, Rizwan Ahmad., IGI Global, USA, (2013).
- [4] F. N. Al-Aboud, "Strategic information system planning: a brief review", *International Journal of Computer Science and Network Security*, vol. 11(5), 2011 pp. 179-183.
- [5] H.B.M. Basir and M.D. Norzaidi, "The effect of strategic alignment on strategic information system planning (sisp) success: an exploratory study in Public Universities in Malaysia", *International Journal of Scientific Research in Education*, vol. 2(2), 2009, pp. 76-87.
- [6] V. Basu, E. Hartono, A.L. Lederer, and V. Sethi, "The impact of organizational commitment, senior management involvement, and team involvement on strategic information systems planning", *Journal of Information and Management*, vol. 39(1), 2001, pp. 513-524.
- [7] T. Bechor, S. Neumann, M. Zviran, and C. Glezer, "A contingency model for estimating success of strategic information systems planning", *Journal of Information and Management "elsevier"*, 2010, pp. 17-29.
- [8] BCG (2013), Growth Share Matrix, accessed by the internet on December (2014), <http://www.businessnewsdaily.com/5693-bcg-matrix.html>.
- [9] T.A. Byrd, B.R. Lewis, and R.V. Bradely, "IS infrastructure: the influence of senior it leadership and strategic information system planning", *Journal of Computer Information System*, 2006, pp. 101-113.

- [10] M.J. Earl, "Approaches to information system planning: experiences in strategic information system planning", *The International Conference on Information Systems*, 1990, pp. 191-225.
- [11] A.C. Hax and N.S. Majluf, "Strategy and the strategy formation process", M.I.T, 1986, pp. 1-22.
- [12] M. Hemmatfar, M. Salehi, and M. Bayat, "Competitive advantage and strategic information systems", *International Journal of Business and Management*, vol. 5(7), 2010, pp. 158-169.
- [13] W.L. Hsu and T.G. Gough, "Information system planning - an integration model", report (2000.13), University of Leeds, School of Computer Studies Research Report Series, 2000.
- [14] IDECO (2011a), Vision and Mission of IDECO, accessed by the internet on December (2014), <http://www.ideco.com.jo/portal/WebForms/VisionAndFuture.aspx>.
- [15] IDECO (2011b), Forbes Study, accessed by the internet on December (2014), <http://www.ideco.com.jo/portal/WebForms/ViewNewsOfNews.aspx?ID=47>.
- [16] IDECO (2011c), E_Bill, accessed by the internet on December (2014), <http://www.ideco.com.jo/portal/WebForms/ViewNews.aspx?ID=164>.
- [17] JTI (2013), Market Share Growth in JTI, accessed by the internet on December (2014), <http://www.jti.com/media/news-releases/market-share-continues-grow-most-key-markets-9-month-earnings-achieve-double-digit-growth/>.
- [18] JTI (2012a), JTI-at-a-Glance, accessed by the internet on September (2014), <http://www.jti.com/our-company/jti-at-a-glance/>.
- [19] JTI (2012 b), Anti- Illicit Trade Compliance Programs, accessed by the internet on December (2014), http://www.jti.com/files/6913/3777/3625/JTI_AIT_FightingContraband_May_2011.pdf.
- [20] JTI (2012c), Core Functions, accessed by the internet on December (2014), <http://www.jti.com/careers/career-areas/it/>.
- [21] JTI (2012d), JTI Innovation, accessed by the internet on December (2014), <http://www.jti.com/our-company/innovation/>.
- [22] JTI (2012e), JTI Brands, accessed by the internet on December (2014), <http://www.jti.com/brands/>.
- [23] H. Kandjani, A. Mohtarami, A.E. Andargoli, and R. Shokoohmand, "A conceptual framework to classify strategic information systems planning methodologies", *Journal of Marine Science ICIES*, vol. (2), 2013, pp. 190-196.
- [24] N. Khani, K.M. Nor, and M. Bahrami, "IS\IT capability and strategic information system planning (SISP) success", *International Management Review*, vol. 7(2), 2011, pp. 75-83.
- [25] K. Koumbati and M. Themistocleous, "Integrating the IT infrastructures in healthcare organizations: a proposition of influential factors", *The Electronic Journal of e-Government*, vol. 4(1), 2006, pp. 27-36.
- [26] A.L. Lederer and V. Sethi, "The implementation of strategic information systems planning mythologies", *MIS Quarterly & the Society for Information Management*, vol. 12(3), 1988, pp. 445-462.
- [27] H.P. Lu, "Managerial behaviors over MIS growth stages", *Management Decision*, vol. 33(7), 1995, pp. 40-46.
- [28] H. Mintzberg, "The fall and rise of strategic planning", *Harvard Business Review*, 1994, pp. 106-114.
- [29] H.E. Newkirk, A.L. Lederer, and C. Srinivasan, "Strategic information system planning: too little or too much?", *Journal of Strategic Information Systems*, 2003, pp. 201-228.
- [30] Z. Pita, F. Cheong, and B. Corbitt, "Major issues in SISP: insights into the main reason of SISP failure", *Journal of European Conference of Information System*, 2009, pp. 182-193.
- [31] Z. Pita, F. Cheong, and B. Corbitt, "Strategic information system planning (SISP): an empirical evaluation of adoption of formal approaches to sisp in Australian Organizations", *International Journal of Strategic Decision Sciences*, vol. 1(2), 2010, pp. 28-61.
- [32] A.J.G. Silvius and J. Stoop, "The relationship between strategic information system planning situational factors, process configuration and success", *Journal of Information Technology & Information Management*, vol. 22(1), 2013, pp. 1-15.
- [33] A. Warr, "Strategic IS planning in UK Organizations: current approaches and their relative success", *Proceedings of the 14th European Conference on Information Systems*, 2006, pp. 972-983.
- [34] Wikipedia (2014), Porter Five Forces Analysis, accessed by the internet on December (2014), http://en.wikipedia.org/wiki/Porter_five_forces_analysis.
- [35] E. Abu-Shanab and L. Bataineh, "Challenges facing e-government projects: how to avoid failure?" *International Journal of Emerging Sciences*, vol. 4(4), 2014, pp. 207-217.

A Novel Web Application for Image Fusion

V. Aslantas, R. Kurban, A.N. Toprak, E. Bendes
Erciyes University, Computer Engineering Department
Kayseri, TURKEY
{aslantas, rkurban, antoprak, ebendes}@erciyes.edu.tr

Abstract— A novel interactive web application for multi focus and multi sensor image fusion is presented in this paper. Basic averaging, Laplacian pyramid, discrete wavelet transform, block-based fusion, spatial domain multi focus image fusion and optimized region based image fusion methods are also included. Users can explore several image fusion methods and compare these methods easily and efficiently by making use of the web application.

Keywords— *image fusion; multi-focus; multi-sensor; web based application*

I. INTRODUCTION

Recently, considering the wide use of imaging technologies, the importance and popularity of image fusion methods has increased. Image fusion has become a significant sub-area of image processing and computer vision [1]. Image fusion is an image processing technique that produces a single synthetic image that contains more complementary information than each of the images of a scene or object in which the images are taken with either more than one sensor or a single sensor with different optical parameters [2].

Fusion of the images obtained by a single sensor with different optical parameters is named as multi-focus image fusion. Optical imaging cameras are seriously affected by the problem of the finite depth of field which means the objects at different distances from the sensor cannot be focused at the same time. As a result, some objects appear in focus (sharp) others defocused (blurred). Multi-focus image fusion aims combining the individual images with different focuses of the same scene or object to gather an everywhere-in focus image [3].

The other image fusion field that included in the proposed image fusion application is multi-sensor image fusion. The aim of the multi-sensor image fusion is obtaining a single composite image by combining the images taken from different sensors [4]. Through this way, obtained fused image is more useful than each of the source images in many areas such as medical imaging, industrial imaging and military applications [5].

Image fusion concepts can be comprehended better by visualizing the fusion processes and results. Thus, image fusion applications should be implemented by visual and interactive and easy accessible web interfaces [6]. Moreover, users have a tendency to explore the theory and applications of the methods by evaluating them with numerous source images. Recently, some interactive applications that includes commercial

software packages, non-commercial toolkits [7, 8] and online HTML based materials, are used to overcome these limitations [9]. However, the non-commercial toolkits have to be downloaded before using. Furthermore, some of them are often require a commercial or academic license. Despite the fact that image fusion related web sites are easy accessible, they are static and do not contain interactive interfaces.

In this paper, a web based image fusion application is presented for multi-focus and multi-sensor image fusion that includes some well-known methods. The main advantages of the proposed application are given as follows:

- Easy accessibility,
- Visualization of the image fusion methods that allow users to comprehend the fusion concepts,
- Comparing the methods by means of fusion performance and robustness,
- Evaluating the methods with differing source images and parameter values.

Furthermore, the users can reach the proposed web based application from everywhere without installing any preliminary other software packages.

The rest of the paper is organized as follows: in the second section, the fusion methods and quality metrics that can be used on web application are explained; in third section, the interactive web based image fusion application is presented; finally, in the last section, conclusions are discussed.

II. IMAGE FUSION METHODS AND QUALITY METRICS

A. Image fusion methods

In this section, image fusion methods available on the web based application are described.

The Laplacian pyramid (LP) consists of filtered and subsampled versions of the original image. The lowest level of the pyramid is the original image. To construct a level of LP, procedures of blurring, reducing image size, interpolating and differencing are employed successively on the previous pyramid level [10]. These procedures are repeated until the envisioned level of the pyramid is reached.

Discrete wavelet transform (DWT) is one of the well-known multi scale decomposition based image fusion methods. To construct the first level of DWT, low-pass and the high-pass filtering followed by down-sampling operations are applied on each row and column of the original image, respectively. By this way, the four DWT sub bands are obtained for the first level of the decomposition. These procedures are recursively repeated on the approximation sub band that contains the horizontal and vertical lower frequencies, until the preferred decomposition level is reached [11].

In the LP and DWT based fusion, first of all, the source images are transformed to a pyramid. Then, the fused coefficient is obtained by selecting maximum coefficients from each corresponding position on the pyramids of the source images. Finally, the fused image is reconstructed by performing an inverse transform.

In block based image fusion (BBIF), first, the source images are divided into equal-sized blocks without spaces or overlaps between the adjacent blocks. Then, for each corresponding block pair, a focus measure is applied to calculate sharpness. The fused image is constructed by copying blocks with the higher sharpness value [12].

Since there is no evidence of how sensitive the fused image to several different values of the block size, a suitable block size needs to be decided. To overcome this problem, the block size is optimized in spatial domain multi focus image fusion (SDMIF) [13].

The optimized region based multi sensor image fusion (ORMSIF) method is based on DWT [14]. Rather than choosing maximum coefficient of DWT, a weighted average of coefficient is utilized. Instead of establishing fusion rules for all the coefficients of the sub bands, the weights are fixed by an optimization algorithm for predetermined regions.

B. Quality metrics

A sharp image contains more dispersed pixel values than the blurred one. Hence, variance of the image is a criterion for measuring the quality of image. For an $M \times N$ size of fused image (f), the variance is computed as:

$$VAR = \frac{1}{M \times N} \sum_i \sum_j (f(i, j) - \bar{f})^2 \tag{1}$$

where \bar{f} is the average grey level over the image region:

$$\bar{f} = \frac{1}{M \times N} \sum_i \sum_j f(i, j) \tag{2}$$

Mutual information (MI) measures the shared information between reference and fused image by using Kullback-Leibler measure as follows [15]:

$$MI_{RF} = \sum_{i,j} P_{RF}(i, j) \log \frac{P_{RF}(i, j)}{P_R(i)P_F(j)} \tag{3}$$

where P_{RF} is the normalized joint gray level histogram of images R and F , P_R and P_F are the normalized marginal histograms of the two images.

Fusion factor (FF) is a metric based on MI aimed to integrate complementary information from multiple sources. For source images (A and B), the quality of fused image (F) is defined by FF as follows:

$$FF = MI_{AF} + MI_{BF} \tag{4}$$

Objective edge based quality (QE) metric is based on Sobel operator edge statistics and calculated as:

$$QE = \frac{\sum_{i=1}^n \sum_{j=1}^m k^a(i, j)w^a(i, j) + k^b(i, j)w^b(i, j)}{\sum_{i=1}^n \sum_{j=1}^m w^a(i, j) + w^b(i, j)} \tag{5}$$

where w^a ve w^b are the edge magnitudes for two source images, k^a ve k^b edge preservation coefficients. For detailed explanation of this metric, see [16].

The $PSNR$, which expressed with regard to the logarithmic decibel scale, is the ratio between a signal's maximum possible power and the power of corrupting noise and given as follows:

$$PSNR = 10 \log_{10} \left(\frac{L^2}{\frac{1}{M \times N} \sum_{i=1}^M \sum_{j=1}^N (R(i, j) - F(i, j))^2} \right) \tag{6}$$

where L is the number of gray levels.

The $SSIM$ is used for calculating the similarity between two images. $SSIM$ is calculated on consequent windows of an image. The similarity between window R and F of a common window size (e.g. 8x8) is:

$$SSIM(R, F) = \frac{(2\mu_R\mu_F + c_1)(2\sigma_{RF} + c_2)}{(\mu_R^2 + \mu_F^2 + c_1)(\sigma_R^2 + \sigma_F^2 + c_2)} \tag{7}$$

where; μ_R is the average of R , μ_F is the average of F , σ^2_R is the variance of R , σ^2_F is the variance of F , σ_{RF} is the covariance of R and F , $c_1=(k_1L)^2$ and $c_2=(k_2L)^2$ are the two variables to stabilize the division, L is the dynamic range (typically 255 for 8-bit gray level images), $k_1=0.01$ and $k_2=0.03$ [17]. $SSIM$ index produces a value between -1 and 1.

III. AN INTERACTIVE WEB BASED IMAGE FUSION APPLICATION

In this section, the proposed interactive web based application is introduced. In order to develop the application two major underlying technologies: Microsoft ASP.NET and MATLAB are used. The reason of the using Microsoft ASP.NET for web interface is that it gives the easy web development skills. MATLAB is used to realize the image fusion tasks, since it presents huge libraries for image processing, wavelet, optimization, and so on.

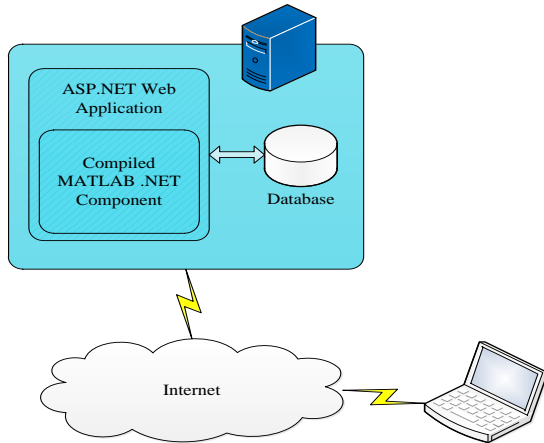


Fig. 1. Architecture of the proposed application

The main purpose of the proposed application is providing an easy to use web page that allows users to access image fusion sources and methods from any place. Users can also change several parameters of the fusion techniques to observe the effects of parameters on the fusion results. The application consists of only one page that acquire source images and the parameter values of the methods from the user and presents visual and quantitative fusion results.

The proposed application is composed of the web interface, the fusion engine and the database. The web interface has a

user-friendly design which allows users to submit the images to be fused, change the parameters of the methods, and get numeric and visual fusion results within minutes. In the design of the web interface, ASP.NET technology (with AJAX extensions) is preferred in order to make the design user-friendly and increase the interactivity of the website. The fusion and evaluation stages are carried out in a compiled MATLAB .NET component which called as fusion engine. That component has implementation of various methods and quality metrics that mentioned in this paper. Fusion engine is developed in MATLAB environment and then embedded into web site. The last component of the application is a MySQL based database that used to save the fusion results. The basic architecture of the proposed application is illustrated in Fig. 1.

In Fig. 2, user data form and parameter selection sections of the main page can be seen. Users are free to decide whether or not to provide their personal information by entering it in the form. Source images need to be uploaded prior to start fusion process. Users can choose the source images from their computer as Windows Bitmap format (.BMP). Users can also reach image database for previously presented multi-focus and multi-sensor test images. According to the uploaded source images, users choose the fusion type as multi-focus or multi-sensor. If the multi-focus is selected as the fusion type, basic averaging, LP, DWT, BBIF and SDMIF methods will be available on the parameter selection section. If the type is selected as multi-sensor, basic averaging, LP, DWT and ORMSIF methods will be available. The parameter sets consist of decomposition level selection and consistency check option for LP, decomposition level and filter family choice for DWT, and block size and focus measure selection and consistency check option for BBIF method. User can select the iteration number, population number, algorithm, crossover rate and mutation rate for the SDMIF method. At last, the parameters of the ORMSIF method are iteration number, population number, algorithm, crossover rate, mutation rate, number of region, and wavelet family. For the SDMIF and ORMSIF methods if the algorithm is selected as ABC, limit parameter will be available instead of mutation and crossover rate parameters.

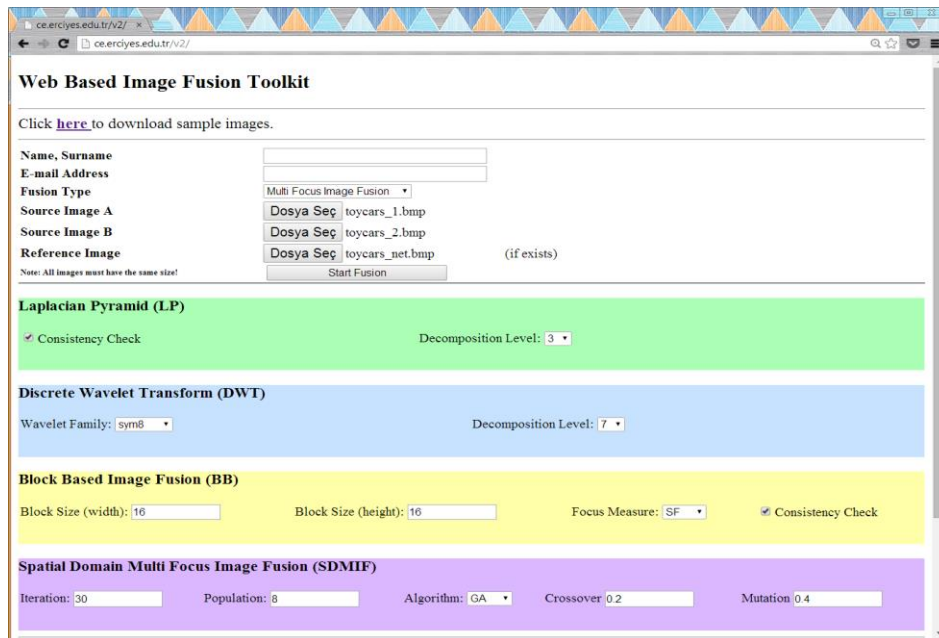


Fig. 2. Main page of web based application. (Parameter Section)

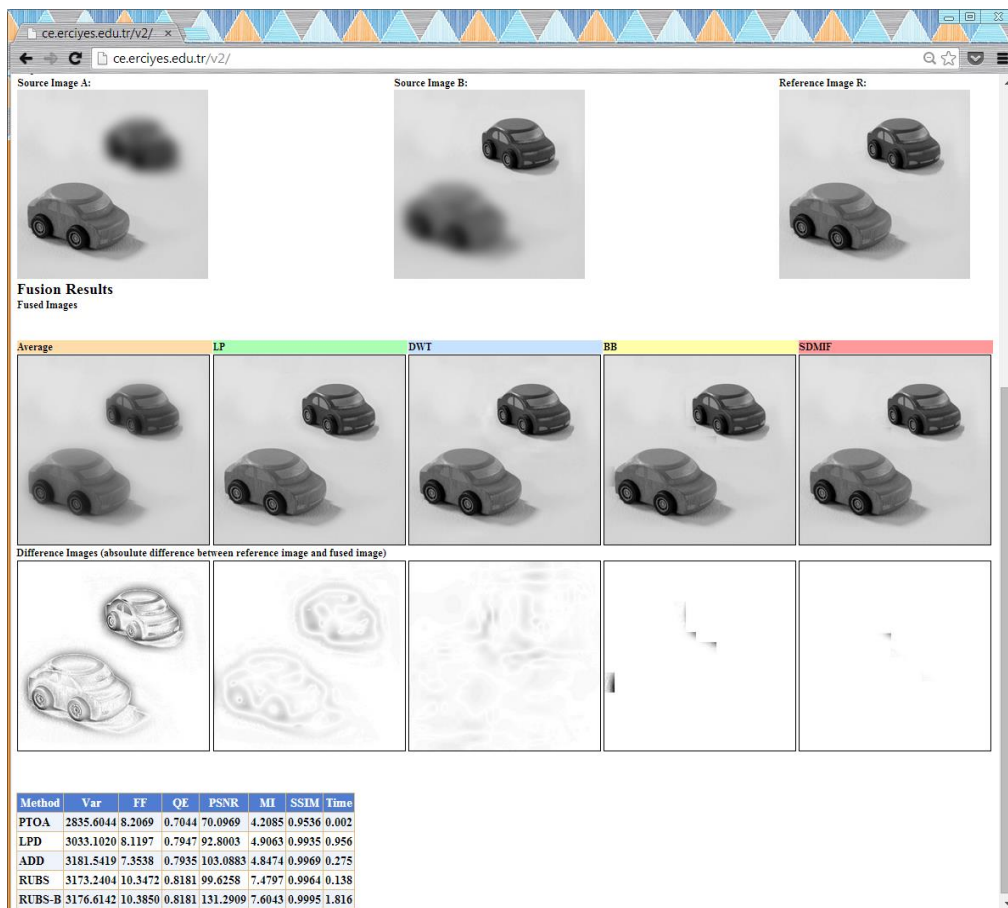


Fig. 3. Main page of web based application. (Parameter Section)

Fig. 3 shows the visual and quantitative results sections of the main page. The source images and the reference image (if exists) that submitted by user are also shown in this section. Fused and difference image of corresponding fusion methods are illustrated in a table. In order to simplify the visual evaluation, the absolute error images which are computed by subtracting the fused image from the reference image are also shown under the fusion results. If there is no reference image, the difference error images are obtained by subtracting the fused image from the each source images.

At last, quantitative results are also given as a table at the bottom of the web page. The table consists of both of the objective metrics and reference based metrics. Included objective metrics are VAR, FF and QE and included reference based metrics are MSE, PSNR, MI and SSIM. Reference based metric results are only available when a reference image is submitted by user.

IV. CONCLUSION

In this paper, a web based interactive application for multi-focus and multi-sensor image fusion that includes well-known image fusion methods is presented.

The proposed application has an easy to use and interactive web interface. It allows the users to visualize the image fusion methods, understand the basic principles of the image fusion, compare the methods by means of performance and robustness and evaluate the methods with different inputs, and parameter values. Moreover, the users can use the web based application from anywhere without installing a preliminary other software packages. The proposed application enables users to conduct experiments on different image fusion methods.

REFERENCES

- [1] V. Aslantas and R. Kurban, "A comparison of criterion functions for fusion of multi-focus noisy images," *Optics Communications*, vol. 282, pp. 3231-3242, Aug 2009.
- [2] V. Aslantas and A. N. Toprak, "A pixel based multi-focus image fusion method," *Optics Communications*, vol. 332, pp. 350-358, 2014.
- [3] Z. B. Wang, Y. D. Ma, and J. Gu, "Multi-focus image fusion using PCNN," *Pattern Recognition*, vol. 43, pp. 2003-2016, Jun 2010.
- [4] V. Aslantas and E. Bendes, "Differential evolution algorithm based spatial multi-sensor image fusion," in *ICINCO 2014 - Proceedings of the 11th International Conference on Informatics in Control, Automation and Robotics*, 2014, pp. 718-725.
- [5] R. S. Blum, Z. Xue, and Z. Zhang, "An Overview of Image Fusion," in *Multi-Sensor Image Fusion and Its Applications*, R. S. Blum and Z. Liu, Eds., ed: Taylor&Francis, 2006, pp. 1-36.
- [6] V. Aslantas, R. Kurban, A. N. Toprak, and E. Bendes, "An interactive web based toolkit for multi focus image fusion," *Journal of Web Engineering*, vol. 14, pp. 117-135, 2015.
- [7] D. Mueller, A. Maeder, and P. O'Shea, "The generalised image fusion toolkit (GIFT)," in *Insight journal, special issue : MICCAI workshop on open science*, 2006, pp. 1-16.
- [8] O. Rockinger. *Image Fusion Toolbox for Matlab*. Available: <http://www.metapix.de/toolbox.html>
- [9] G. Bebis, D. Egbert, and M. Shah, "Review of computer vision education," *IEEE Transactions on Education*, vol. 46, pp. 2-21, Feb 2003.
- [10] R. Blum, Z. Xue, and Z. Zhang, *An Overview of Image Fusion*: CRC Press, 2005.
- [11] G. Pajares and J. M. de la Cruz, "A wavelet-based image fusion tutorial," *Pattern Recognition*, vol. 37, pp. 1855-1872, Sep 2004.
- [12] S. Li, J. T. Kwok, and Y. Wang, "Combination of images with diverse focuses using the spatial frequency," *Information Fusion*, vol. 2, pp. 169-176, 2001.
- [13] V. Aslantas and R. Kurban, "Fusion of multi-focus images using differential evolution algorithm," *Expert Systems with Applications*, vol. 37, pp. 8861-8870, 2010.
- [14] V. Aslantas, E. Bendes, R. Kurban, and A. N. Toprak, "New optimised region-based multi-scale image fusion method for thermal and visible images," *IET Image Processing*, vol. 8, pp. 289-299, 2014.
- [15] Q. Guihong, Z. Dali, and Y. Pingfan, "Information measure for performance of image fusion," *Electronics Letters*, vol. 38, pp. 313-315, 2002.
- [16] C. S. Xydeas and V. Petrovid, "Objective image fusion performance measure," *Electronics Letters* vol. 36, pp. 308-309, 2000.
- [17] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, pp. 600-612, Apr 2004.

Ontology-based Facilitation Support Tool for Group Decision Making

Abdelkader Adla, Bakhta Nachet

Department of Computer Science
University of Oran 1 Ahmed Ben Bella
Oran, Algeria
{adla.abdelkader, nachet bakhta}@univ-oran.dz

Abstract—The need for Group Decision Making (GDM) techniques and support is greater than ever before. In the group decision making process, the alternatives amongst which a decision must be made can range from a few to a few thousand; the facilitator (or the decision makers) need(s) to narrow the possibilities down to a reasonable number, and categorize and classify alternatives. Even when this is not the case, facilitation support, such as ontology-based frameworks potentially offer these capabilities and can assist the decision-maker in presenting the alternatives in a form that facilitates the decision. In this research an ontology base approach is developed to facilitate organizing alternatives. The resulting alternatives organizing tool is based on two ontologies. These two ontologies are supplementary and each one ensures an aspect of the decision organizing. They have been built using the Web Ontology Language (OWL) which facilitates the sharing and integration of decision-making information between multiple decision makers.

Keywords—GDSS, ontology of the domain of application, ontology of domain, organization of alternatives, OWL

I. INTRODUCTION

The need for Group Decision Making (GDM) techniques and support is greater than ever before. This is due to the complexity of business relationships, the greater number of decision makers and organizations that are involved in the decision process, online access to multiple external information sources, and the decreasing in the time allowed for decision making.

In the group decision making process, the alternatives amongst which a decision must be made can range from a few to a few thousand [1][2]. The facilitator (or the decision makers) need(s) to narrow the possibilities down to a reasonable number, and categorize and classify alternatives, especially where the alternatives can be put into numerical terms. Even when this is not the case, facilitation support, such as ontology-based frameworks potentially offer these capabilities and can assist the decision-maker in presenting the alternatives in a form that facilitates the decision.

In this research, an ontology based approach is developed to facilitate organizing alternatives during the group decision making process. The alternatives organizing tool is based on two ontologies: application-domain ontology and domain ontology.

The first ontology will allow structuring all documented possible decisions by specifying semantic inter-relations. The domain ontology defines the objects of the domain as well as their inter-relations. This second ontology will ensure another aspect of the generalization link between decisions. As a result, these two ontologies are

supplementary and each one ensures an aspect of the decision organizing.

We have built the ontologies using the Web Ontology Language (OWL) which facilitates the sharing and integration of decision-making information between multiple decision makers via the Web and Description Logic.

The remainder of this article is structured as follows: a background on group decision making and decision support is given in section 2. The section 3 presents related works. In the section 4, we develop our ontology-based approach to organize alternatives decision in the group decision making process. Section 5 is devoted to a detailed presentation of the developed ontologies to facilitate the alternatives organizing, followed par an illustration with an example in section 6. Finally, in section 7 we conclude and give future work.

II. BACKGROUND

Decision aid and decision making have greatly changed with the emergence of information and communication technology (ICT). Decision makers are now far less statically located; on the contrary they play the role in a distributed way. This fundamental methodological change creates a new set of requirements: distributed group decision making is necessarily based on incomplete data, it must be possible at any moment, and it might be necessary to interrupt a decision process and to provide another, more viable decision. “Distributed group decision making” means that several entities (humans and machines) cooperate to

reach an acceptable decision, and that these entities are distributed and possibly mobile along networks [1].

In [1], the authors consider the paradigm of distributed group decision-support systems, in which several decision-makers who deal with partial, uncertain, and possibly exclusive information must reach a common decision. To this end, the use of a group system makes possible the collaboration of distant decision makers. The cooperative work so initiated can be synchronous or asynchronous. A small group or a whole organization can be supported. The application can be carried in several sites over a common information base. The networked decision makers work together to solve a particular problem although they might neither be present at the same time in the same place nor constitute a permanent organization. Thus, decision-makers can evaluate and rank alternatives, determine the implications of offers, maintain negotiation records, and concentrate on issues instead of personalities.

Experience with group decision making has shown that an on-line “meeting” is generally used to represent a group decision process for the specific problem at hand and a recurring pattern of three stages occurs in the group decision process [1]. These three process phases are: Pre-meeting, during meeting, and post-meeting (Fig. 1).

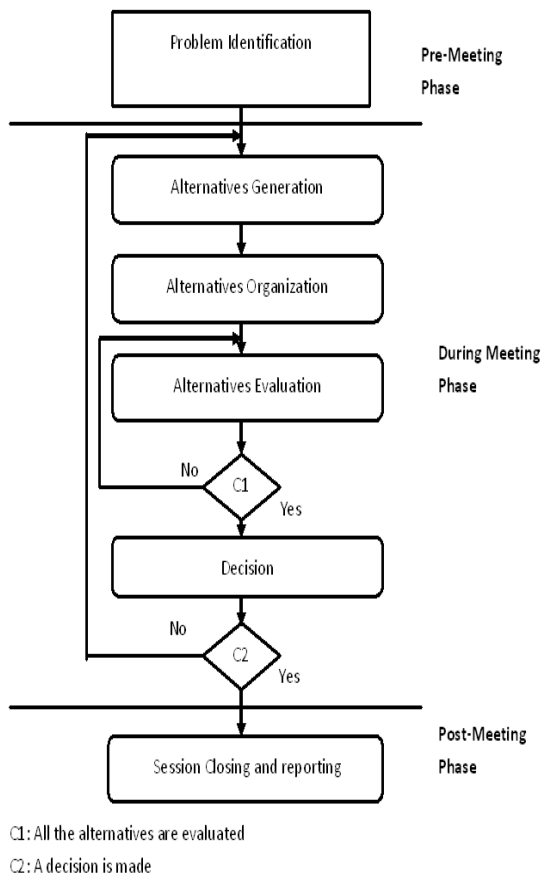


Figure 1 : Group Decision Making Process Model [ADLA 10]

The group decision model assumes that decision-makers are located in different places. A computer network is presumed that connects these different locations of participants. The decision-making process is controlled by a facilitator. The facilitator initiates, prepares the phases of the decision making process. He defines the issue(s) for decision and organizes the human group of decision makers for the decision-making process. His responsibility is to distribute the results among the participants after the decision-making. During the process, the mediator has a principal responsibility for the convergence of decision making process. He is responsible for the complete process and its deliverable, namely the decision.

In “during meeting” phase, a group can generate many alternatives in a short period of time. These alternatives may be similar or duplicated that need to be merged. The redundant alternatives can be retrieved for the facilitator to review, and then they can be merged or deleted. Idea organization in a distributed environment is mainly the facilitator’s responsibility. It can be a very challenging task for the facilitator.

III. RELATED WORK

Ontologies are used as part of the improvement of the management of an organizational memory. In this perspective, ontology is mainly used to manage large case bases by facilitating their storage, representation and information semantic retrieval. Among the systems which use this aspect of ontologies in DSS, we cite the platform PROTEUS [3]. In the same context, the tool TextViz [4] is used information semantic retrieval in the field of car diagnosis. The ontology in that case, represents the knowledge concerning the breakdowns diagnosis. In [5], the authors propose an ontology of gas turbine and a reasoning tool based on that ontology

In [6], the author develops an ontology- based system to support the risks analysis in industrial domain. This system used resources indexing and a case base reasoning. Also in [7], the authors developed domain ontology to assist jurists during the juridical problem solving.

Overall the research in [8] provides the first attempt at documenting, storing, and retrieving engineering design decisions using ontologies and provides the foundation for the development of a more comprehensive decision support framework.

IV. THE ONTOLOGY BASED APPROACH FOR ALTERNATIVES ORGANIZING IN GROUP DECISION MAKING

The alternatives proposed by the decision-makers can contain decisions which are:

- Redundant: the alternatives are syntactically identical;
- Synonyms: the alternatives are syntactically different, but semantically identical;

- **Conflicting:** two contradictory or conflicting alternatives mean that the application of one is incompatible with the application of the other;
- **Generic:** an alternative may be more general than another. In this case, the application of the most general includes the application of the most specific;

These alternatives must be organized before being evaluated and thus enabling the decision choice. Our work consists to organize these alternatives. The alternative organizing contributes to retrieve and remove all the redundant, conflicting and synonymous decisions. Besides, when an alternative is more general than another, both the alternatives are presented to the decision-makers and it is their duty to choose one.

The main role of the organizing tool is to allow identifying semantic relationships between decisions then to present them to the decision-makers who will have the duty to decide among the suggested alternatives which will be removed and which have to be kept based on their expertise and the semantic relationships existing between the generated decisions.

V. THE PROPOSED ONTOLOGIES

The purpose of our work is to integrate an organizing tool (Fig. 2) into a Group Decision Support System (GDSS) to support the facilitator during the alternatives organizing stage in the group decision making process. Our ontology based approach to support alternatives organizing uses two ontologies:

1) *Application domain ontology:* It will be used in the alternatives organizing by the group of the decision-makers. It is a conceptual ontology where each object represents an alternative decision proposed by a decision-maker as a solution to the breakdown diagnosis problem. The application domain ontology specifies all decisions and the relations between them. Indeed, two decisions which can seem at first glance semantically close can be contradictory or incompatible in the context of the diagnosis of breakdowns application. This is why it is necessary to consider relations between decisions according to their effects on a particular task, i.e. the equipment maintenance, and not analyze a decision upon its syntactic expression based on the domain ontology.

2) *Domain ontology:* This ontology specifies concepts which are the equipment components. Relations between these concepts are of aggregation and inclusion. In effect a component which is included (directly or indirectly) in another is linked to the latter by a semantic relation of generalization. Domain ontology of the equipment concerns the vocabulary used in the expression of the decisions in terms of equipment components. The domain ontology is considered to be an explicit specification of concepts relating to the equipment maintenance as well as the relations existing between these concepts.

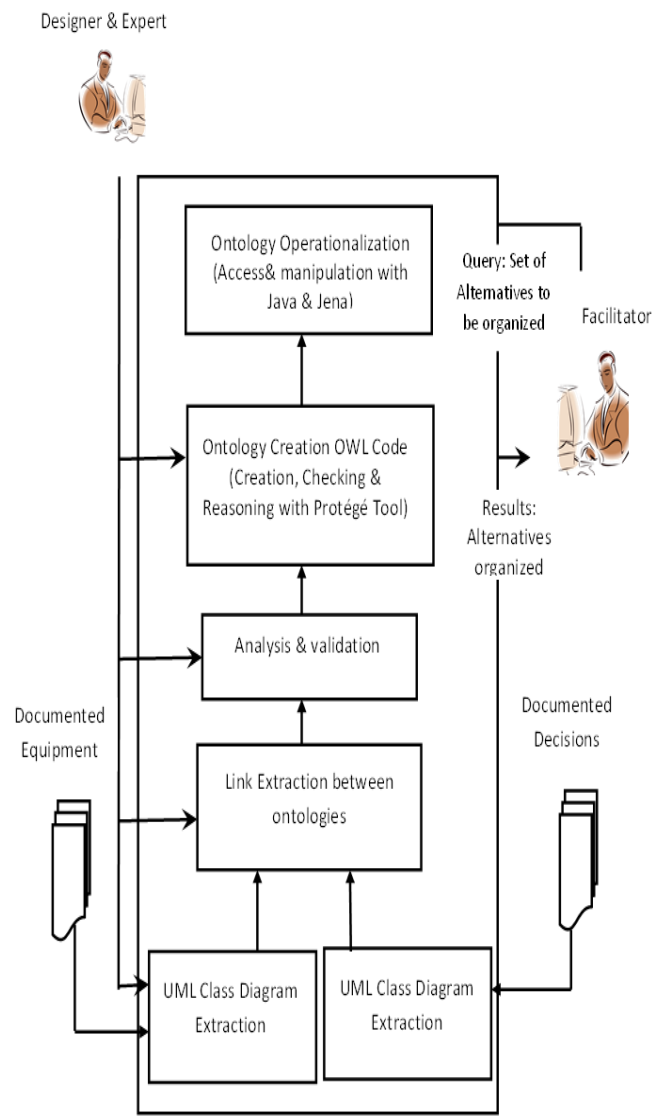


Figure 2: Functional Architecture of the Alternatives organizing tool

The link between both ontologies is materialized by the fact that in the application domain ontology each decision is indexed by one or several objects (components) implied by this decision. The use of two distinct but smoothly coupled ontologies will enable to infer new useful knowledge for the alternatives organizing task. Both ontologies must be fully developed. The general approach cited in [9] is adopted to develop both ontologies (Fig. 3). The three stages of the approach (conceptualization, ontologization and operationalization) are in general preceded of requirement analysis and knowledge domain delimitation. This process must however be entirely validated by a human expert.

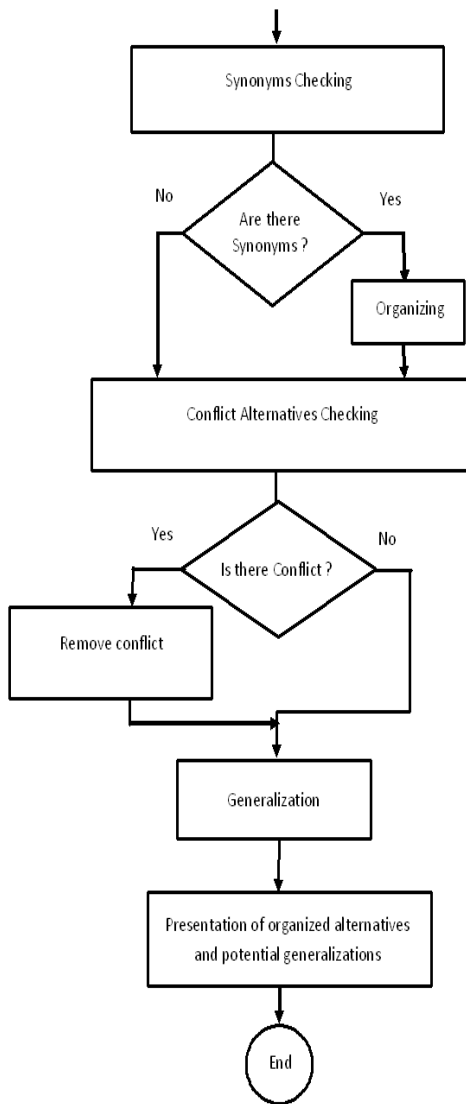


Figure 3: General functioning diagram of the alternatives organizing tool

We consider the breakdowns diagnosis application in a complex industrial system. In this kind of systems, decisions are known and listed in an appropriate documentation. The decision-makers who are experts in their domains propose decisions as possible solutions to the problem. Faced with the huge amount of alternatives decisions suggested by the decision makers, the facilitator has to come to a consensual decision. The integration of an organizing tool in the GDSS is the first stage in preparing the decision choice. In this regard, the tool is useful and will give a significant support to the facilitator.

A. Conceptualization

This stage consists in representing ontology by a conceptual model in a high level of abstraction. The used conceptual model represents the concept classes and their instances. We use UML class diagram to represent the conceptual models of ontology (Fig. 4) (Fig. 5). The ontology models allow representing domain concepts of classes and relations between the classes. Every concept or instance may be identified by URI. These models will be of use as inputs of the ontologization stage.

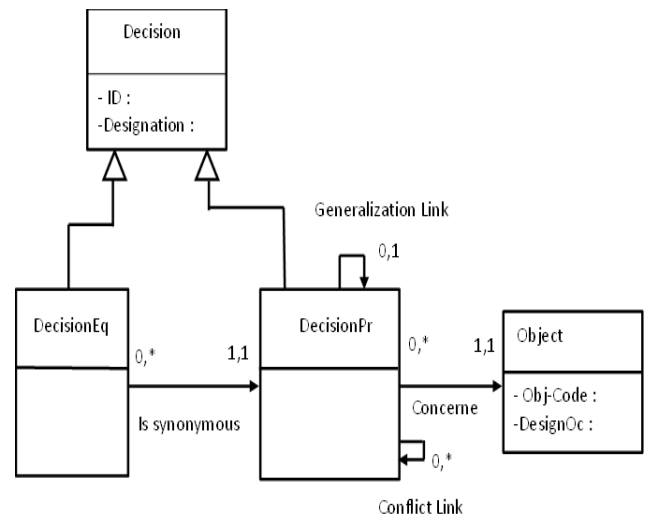


Figure 4: Conceptual Model of the Application domain Ontology

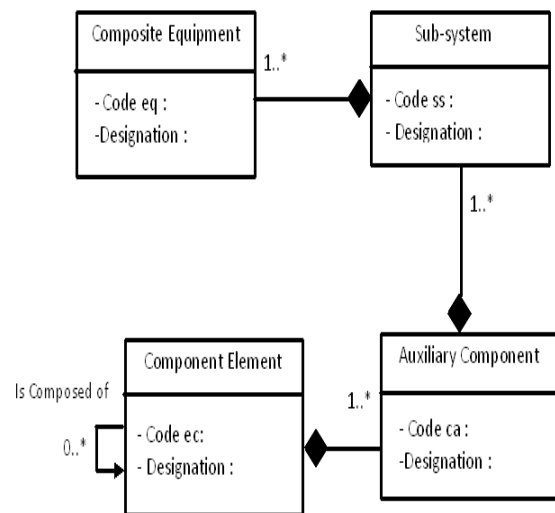


Figure 5: Conceptual Model of the domain ontology

1) *Application domain ontology*: We define four classes:

a) *Decision class*: represents decisions;

b) *DecisionEq class*: represents equivalent decisions; it is a sub-class of the decision class. Several equivalent decisions can be related by a synonymy link with a main decision. These equivalent decisions are all different expressing forms of the same decision. This group of decisions is represented by one decision in the main decisions class which contains no semantic redundancy;

c) *DecisionPr class*: represents all main decisions which are interrelated semantically; it is a sub-class of the decision class;

d) *Objetconcerned*: represents objects (components) concerned by decisions.

We define three types of relations:

a) *Conflict relations*: links a main decision with all the main decisions which are conflicting. A main decision can have several contradiction links. This relation is symmetrical

b) *Generalization relation*: links a decision with its generic decisions. This relation is used to identify a first form of generalization between decisions inferred from the application domain ontology. Example: in the equipment maintenance, replacing a component is more general than repairing it. This relation is transitive;

c) *Synonymy relation*: a group of equivalent decisions is represented by one main decision. So, this relation will be used to identify synonymies between decisions. This relation is functional.

2) *Domain ontology*: We define four classes:

a) *Composite equipment class*: represents a composite equipment to maintain

b) *Sub-system class*: represents the sub-systems which compose an equipment;

c) *Auxiliary component class*: represents all auxiliary components which compose the sub-systems;

d) *Component element class*: represents the elements which compose the auxiliary components.

As for the semantic relations, we define four ones:

a) *"Is composed of" relation*: links the instances of Component element class. This link expresses the relation of composition of a component in another. This relation is transitive; it is used to identify a second form of generalization between decisions inferred from the domain ontology of the equipment to maintain and uses the relation « is composed of ».

b) *Aggregation type relations*: The composite equipment is formed of a group of sub-systems; each of which is formed in return of a group of auxiliary components. Each of the latter contains a group of elements components.

B. Ontologization

Ontologization consists of formalizing conceptual models developed at the previous stage, as far as possible. We use OWL (Ontology Web Language) [10] as formalizing language of our ontologies. OWL is a developing information technology of the Semantic Web and is based in Description Logic (DL) [11]. Description logic is a family of knowledge representation languages used to formally represent knowledge of a domain in a structured manner. OWL represents ontology by building hierarchies of classes which describe the concepts in a domain and the properties which relate these classes to each other.

The creation of our ontologies is done using Protégé Ontology Editor which is an ontology development tool developed by Stanford Medical Informatics. [12]. This allowed the classes and properties to be easily created in an OWL-DL representation. We have also used it to check on our ontologies and the inconsistencies thanks to the reasoner FACT++. It allows as well inferring new knowledge from semantic relations. Then, we generated our ontologies in OWL format.

Example of the individual decision "change_the_parvex-variator" :

```
<owl:NamedIndividual
rdf:about="http://www.ontoproject.org/ontologydecision#change_the_parvex-variator">
  <rdf:type rdf:resource="http://www.ontoproject.org/ontologydecision#Prdecision"/>
  <ID rdf:datatype="&xsd:int">19</ID>
  <Designation rdf:datatype="&xsd:string">the variator is faulty, it must be replaced</Designation>
  <general
rdf:resource="http://www.ontoproject.org/ontologydecision#check_the_connection_of_the_variator_plug"/>
  <concern rdf:resource="http://www.ontoproject.org/ontologydecision#parvex_variator"/>
</owl:NamedIndividual>
```

Fig.6 depicts partial view of the application domain ontology. The URI base is:"<http://www.ontoproject.org/ontologydecision>".

Example of the individual Component element "parvex-variator"

```
<owl:NamedIndividual
rdf:about="http://www.ontoproject.org/ontologyequipment#parvex-variator">
  <rdf:type
rdf:resource="http://www.ontoproject.org/ontologyequipment#Componentelement"/>
  <Codeec rdf:datatype="&xsd:int">7</Codeec>
  <Designation rdf:datatype="&xsd:string">searching a label for parvex variator</Designation>
```



```

    <iscomposed
    rdf:resource="http://www.ontoproject.org/ontologyequ
    ipment#CAN-network"/>
    <iscomposed
    rdf:resource="http://www.ontoproject.org/ontologyequ
    ipment#axial-variator"/>
    <iscomposed
    rdf:resource="http://www.ontoproject.org/ontologyequ
    ipment#brushless-motor"/>
    <iscomposed
    rdf:resource="http://www.ontoproject.org/ontologyequ
    ipment#communication-network"/>
    <iscomposed
    rdf:resource="http://www.ontoproject.org/ontologyequ
    ipment#internal_fuse"/>
    </owl:NamedIndividual>
    
```

Fig. 7 depicts partial view of the equipment domain ontology. The URI base is: <http://www.ontoproject.org/ontologycomponents>.

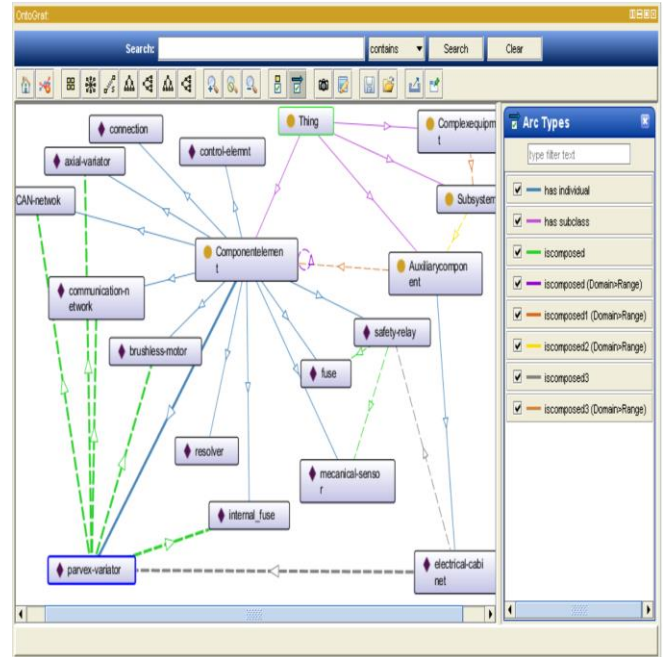


Fig. 7. Partial view of the equipment domain ontology

C. Operationalization:

The figure 8 represents the functioning sequence diagram of the proposed organizing module. Ontology 1 is the application domain ontology while ontology 2 represents the equipment domain ontology.

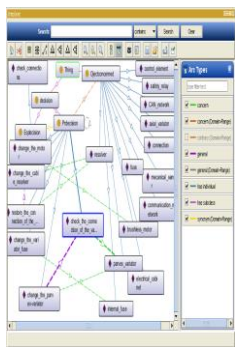
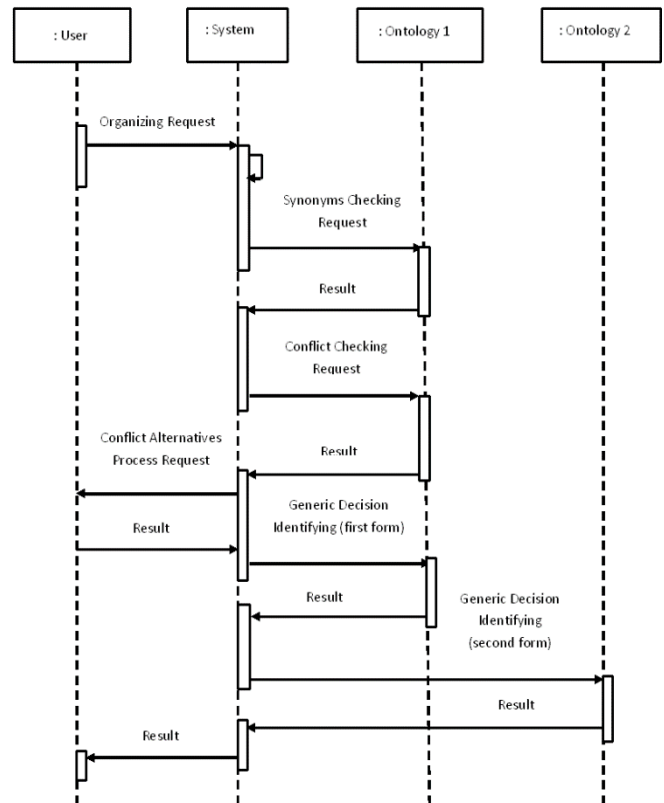


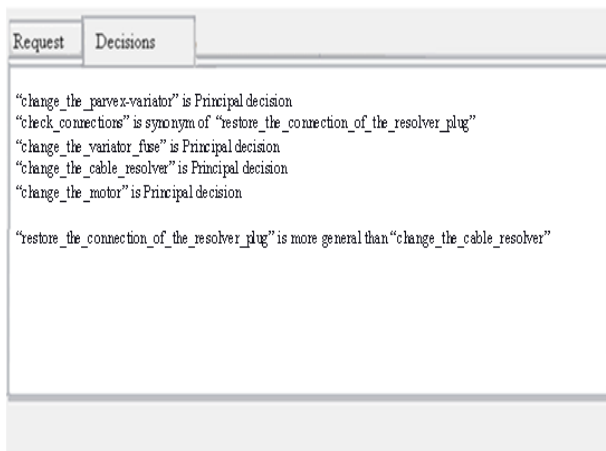
Fig. 6. Partial view of the application domain ontology.

Fig. 8: Sequence diagram of the alternatives organizing step

To operationalize ontologies, we used NetBeans development environment and Java language. To exploit the ontologies, we used Jena framework [13], [14] which provides a programming environment for RDF, RDFS [15] and OWL as well as a querying engine to execute SPARQL queries (Simple Protocol And RDF Query Language) [16] which is RDF querying language.

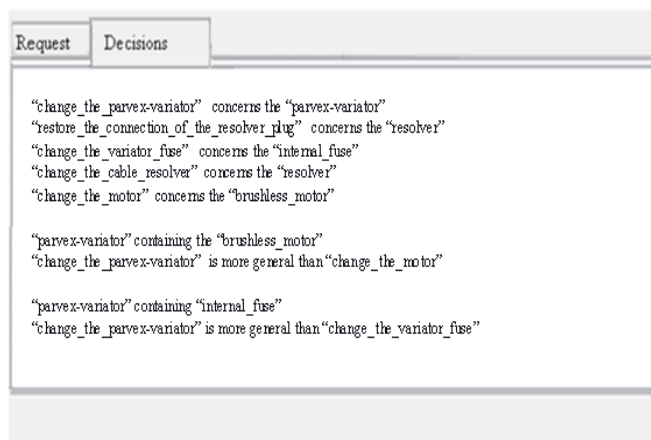
VI. EXAMPLE

Given the set of alternative decisions generated by the group of decision makers and introduced as input to the alternative organizing tool. The latter will process these alternatives in two stages: the first one involves the application domain ontology. The outputs of this stage are synonymous and conflict alternatives as well as the first generalization form.



When two alternatives are conflicting, the facilitator has to remove one. For instance, the decision "restore_the_connection_of_the_resolver_plug" is conflicting with "change_the_cable_resolver", the facilitator has chosen to remove the decision "change_the_cable_resolver". Thus, the latter don't appear in the following stage.

The result of this organizing stage is as follows:



VII. CONCLUSION

In this paper we presented an ontology-based approach for to support facilitation in the group decision making. The alternatives organizing facilitation tool supports the facilitator in the group decision making process. To this end, we have developed two ontologies: application domain ontology and domain ontology. The first one is relating to alternatives organizing task. It structures decisions and their semantic relationships. As domain semantics are not entirely expressed by this ontology, the latter is connected to a second ontology which supplements semantics by specifying the knowledge of the domain upon which decisions are applied.

The jointly use of both ontologies allows organizing and categorizing the alternatives decisions.

As future work, we plan to extend our approach by developing a third ontology: task ontology. This latter will express the context of the problem solving task.

REFERENCES

- [1] A. Adla, P. Zarate and J.L. Soubie, "A Proposal of ToolKit for GDSS Facilitators, Group Decision and Negotiation (GDN), vol. 1, 2011, Springer
- [2] G. Marakas, Decision Support Systems in the 21st Century, 2nd Edition, Prentice Hall, 2003.
- [3] I. Rasovsca, B. Chebel Morello and N. Zerhouni, "A case elaboration methodology for a diagnostic and repair help system based on CBR", in "20th International Florida Artificial Intelligence Research Society Conference (FLAIRS'07), Key West, Floride, USA, 2007.
- [4] A. Reymonet and J. Thomas, "Ontologies et Recherche d'Information : une application au diagnostic automobile", 21èmes Journées Francophones d'Ingénierie des Connaissances, Nîmes France, 2010.
- [5] F.Z. Laallam and M. Sellami, "Gas turbine ontology for the industrial processes", Journal of Computer Science vol. 3 (2), 2007, pp. 113-118.
- [6] B. Debray, C. Duval, A. Jovanovic and O. Salvi, "Integrated management of emerging risks: challenges and objectives of the iNTeg-Risk European project", in 16ème Congrès Lambda-Mu, Avignon, France, 2008.
- [7] K. Dhoubi, S. Despres and F. Gargouri, "Structuration des décisions de jurisprudence basée sur une ontologie juridique en langue arabe", in 12e Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC), 2012, pp. 303-308.
- [8] G.M. Mocko, D.W. Rosen and F. Mistree, "Decision Retrieval and Storage Enabled Through Description Logic", in ASME IDTC /CIE, 2007.
- [9] F. Furst, "Contribution à l'ingénierie des ontologies : une méthode et un outil d'opérationnalisation", Thèse de doctorat, Université de Nantes, 2004.
- [10] <http://www.w3.org/TR/owl-features/>
- [11] T. Berners-Lee, J. Hendler and O. Lassila, "The Semantic Web", Scientific American Magazine, 2001.
- [12] http://protege.stanford.edu/doc/users_guide/index.html.
- [13] <http://jena.sourceforge.net/>
- [14] <http://www.hpl.hp.com/semweb/jena.htm>.
- [15] <http://www.w3.org/TR/rdf-schema/>
- [16] <http://www.w3.org/TR/rdf-sparql-query/>

Quality Driven Approach for Data Integration Systems

Mohamed Samir Abdel-Moneim

College of Computing and Information Technology
Arab Academy for Science Technology & Maritime Transport
Cairo, Egypt
moh_samir_86@hotmail.com

Ali Hamed El-Bastawissy

Faculty of Computer Science
MSA University
Giza, Egypt
aebastawissy@msa.eun.eg

Mohamed Hamed Kholief

College of Computing and Information Technology
Arab Academy for Science Technology & Maritime Transport
Alexandria, Egypt
kholief@aast.edu

Abstract—By data integration systems (DIS) we mean the systems in which query answers are instantaneously mapped from a set of available data sources. The query answers may be improved by detecting the quality of the data sources and map answers from the significant ones only. The quality measures of the data in the data sources may help in determining the significant data sources for a given query. In this paper, we suggest a method to calculate and store a set of quality measures on data sources. The quality measures are, then, interactively used in selecting the most significant candidates of data sources to answer user queries. User queries may include the user preferences of quality issues. Quality-based approach becomes increasingly important in case of big number of data sources or when the user requires data with specific quality preferences.

Keywords—*data integration; quality measures; data sources; query answers; user preferences*

I. INTRODUCTION

Data Integration (DI) is the process of finding and retrieving data located at multiple locations, and allowing the user to view these data through a single unified view called global or mediated schema [1, 2]. Users use the global schema to pose their queries to a data integration system. The global schema provides a uniform access to data stored in heterogeneous and autonomous sources. The user no longer needs to consider which sources are relevant to their queries, how the data are structured at the sources, how to access the data sources, nor does he need to consider how to combine the results from different sources. Data integration system (DIS) queries the data sources according to the location of the required data:

- The required data may be found at a single source. In this case, the system has to query only that particular source.
- The required data may be found at many sources. In this case, the system may choose to query multiple sources or to query the best sources and combine the results.

- The required data may be scattered across many sources. In this case, the system must query different sources and combine the result from each source.

Different architectures for data integration systems have been proposed, but broadly speaking, most systems fall between warehousing and virtual integration [3]. In the data warehouse system, data from different homogeneous or heterogeneous sources are loaded into a physical database called warehouse through a process called extract, transform and load (ETL) so that queries over the data warehouse can be answered as shown in “Fig. 1”.

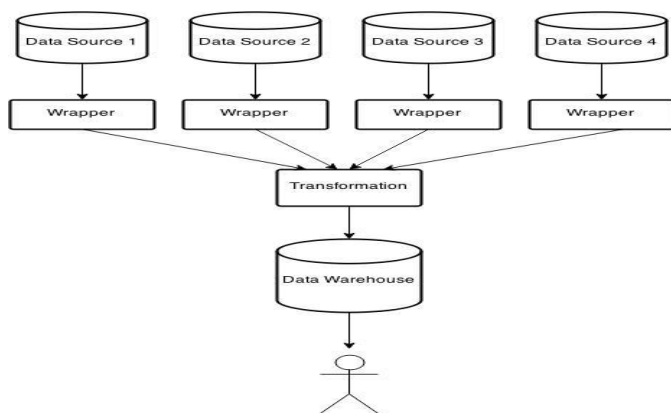


Fig. 1. ETL process

In virtual integration system, data remain in the sources and are accessed as needed at the query time. Data integration system (DIS) is often built as a mediator-wrapper architecture [4] as shown in “Fig. 2”. Although the two approaches are different, many problems are shared across their architectures.

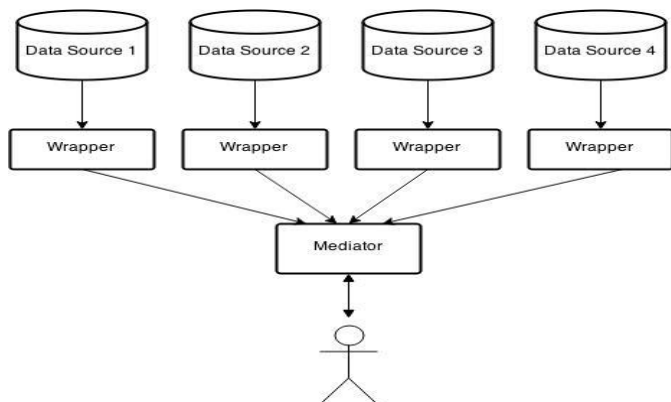


Fig. 2. An architecture of a data integration system

The quality of the data sources can dramatically change as data may be incomplete, inaccurate or out of date. In fact, the quality of the result depends mainly on two factors: the quality of the data at the data sources and the manipulation process that generates the resulting data from the data sources. Because of the high number and high diversity of participating data sources as well as their autonomy, it is important to store some quality-related measures to take it into consideration during query planning. Data Quality (DQ) has become very important in organizations and many application domains [5, 6]. DQ is based on a set of dimensions such as timeliness, completeness, and accuracy.

In this paper we present an approach that incorporates data quality into data integration systems in order to get satisfactory query plans. Our approach is based on adding quality system components such as data quality acquisition to be parts of any data integration system. We integrate attribute values from different data sources based on quality measures and user’s preferences. We use quality measures to deliver query answers with satisfactory quality.

The rest of this paper is organized as follows. In Section II, we briefly discuss data quality dimensions for data Integration. Section III describes the work related to quality based data integration systems. We illustrate the architecture and functions of our data integration quality system components in section IV. Section V describes our quality driven query processing algorithm. We evaluate and validate our approach in section VI. The conclusion and future work are presented in Section VII.

II. DATA QUALITY CRITERIA FOR DATA INTEGRATION

Broadly defined, data quality means “fitness for use” [7, 8]. Therefore, the interpretation of the quality of data item depends on the user’s needs. While some data quality may be appropriate for a given task or user, it may not be appropriate for another user or another task. Data quality dimensions depend on each other and only a suitable set of dimensions is appropriate for a given task. To decide which data quality dimensions to use, Wang and Strong [9] have empirically defined fifteen data quality dimensions considered by end users as the most significant. Wang and Strong classify these dimensions into contextual, intrinsic, representational and accessibility quality as shown in “Fig. 3”.

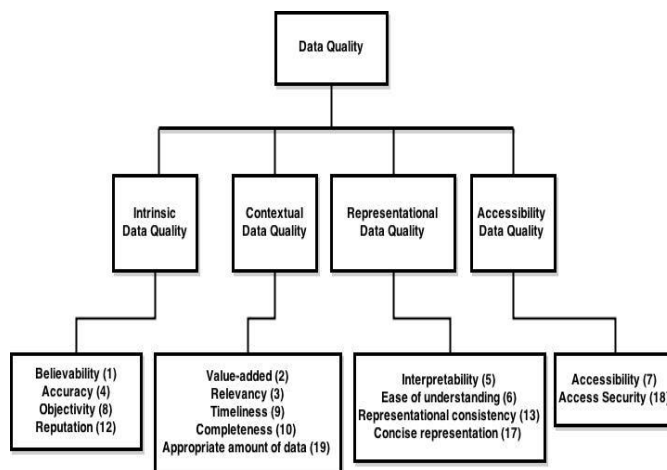


Fig. 3. A conceptual framework of data quality

The measurements of data quality dimensions can be done at different granularities:

- 1) Data source level: determine the quality for the whole source. Quality measures of this type remain unchanged as long as the source doesn’t dramatically change.
- 2) Relation level: determine the quality of a relation in a data source.
- 3) Attribute level: determine the quality of an attribute in a relation.

In this we only focus on data quality dimensions that could affect the data integration process and could be considered important from user’s perspective. We illustrate these dimensions as follows:

A. Accuracy

Several definitions have been defined for the term accuracy. Wang and Strong [9] defines accuracy as “The extent to which data are correct, reliable, and certified free of error”. Redman [10] defines accuracy as “the degree of correctness of a given collection of data”. In general, two types of accuracy are considered important by literature, syntactic and semantic [11]. Increasing accuracy of the query answer is important from user’s prospective as data sources might contain incorrect or misspelling data.

B. Completeness

Wang and Strong define completeness as “the extent to which data are of sufficient breadth, depth, and scope for the task at hand” [9]. One of the main objectives of integration is to increase completeness. Completeness is one of the most important data quality dimensions in the integration of data sources. Querying one data source gives a set of results. Querying another data source gives another overlapping set. As the number of data sources queried increase, the result will be more complete.

C. Cost

Cost is the amount of money required for a query. Some data sources may charge users for accessing their data. Considering cost is important so that users can choose between free and commercial data sources.

D. Response Time

It is the amount of time when the mediator submit a query and receive the complete response from the data source. Users usually prefer data sources that have low response time. Response time is important in order to determine the time-outs and unavailability of data sources. Users waiting a long time for a response are more willing to terminate the query. Response time also could be one of the factors for source selection when a data integration system decides which data sources to query in order to answer a query.

E. Timeliness

Timeliness is how old the data are in a data source [12]. Timeliness in the context of data integration is the time between the last verification or update of the data and now. Timeliness is important as some data sources might be outdated and the user might be interested in getting up-to-date data.

III. QUALITY BASED DATA INTEGRATION SYSTEMS

In this section, we present an overview of research projects that have been proposed to perform query processing based on data quality. We focus on how they measure and store data quality, how they process queries and user interference option.

A. The DaQuinCIS architecture

The DaQuinCIS project [13] is designed to improve data quality in cooperative environments.

DaQuinCIS uses metadata to store the quality measures, the interpretation of the quality measures, and information related to the measurements.

DaQuinCIS follows global-as-view (GAV) approach for processing queries. DaQuinCIS decomposes queries submitted over a global schema to queries against local data sources. The query processing approach adopted by DaQuinCIS to find an answer to a query is structured as following:

- 1) A user posed a query Q on the global schema.
- 2) The query Q is then decomposed according to the schema mapping that maps each concept of the global schema in terms of the local sources. Therefore, the query Q is unfolded to Q1,...Qk queries to be sent over the local data sources.
- 3) Executing queries Q1,...Qk, returns results R1,...Rk. A record matching algorithm is used to find items common to both results.
- 4) The final result is built according to the following rules:
 - a) If no quality constraint is specified, the result is generated by selecting the best quality values.
 - b) Whether there are quality constraints, the result is generated by examining whether the constraints satisfy the whole result.

B. Data Integration Techniques based on Data Quality Aspects

Gertz and Schmitt [14] are used data quality to develop data integration techniques within an object oriented data model and used a metadata to store the information about data quality. Quality dimensions such as accuracy, completeness and timeliness are selected for the purpose of database integration. Gertz and Schmitt have also developed query language extensions to be used for specifying data quality goals for global queries and in data integration.

If objects are conflicted semantically, the object with the best data quality must be chosen. If conflicts exist between the integrated objects but they are different at their quality level, then these objects need to be grouped in order to rank the results.

Regarding user contributions in the integration process, the user has less flexibility in determining priorities of the quality dimensions. Because the data quality are offered as the most up to date or the most accurate and not offered in weights or percentages. Consequently, users will not be satisfied by combinations of quality priorities. One result might satisfy a user for a particular task, but of poor quality for other. Also, the user has no option if he wants to integrate more than data source to find a more complete result. Gertz and Schmitt propose the extended query language to deal with the query which take into account the quality feature.

```
Select [restrict] < list of attribute >  
from G  
where < selection condition >  
using < selection feature >  
with < weight feature >
```

The “where” clause applies a condition on the answer set while the “using” clause applies a feature condition on the result set.

C. Quality-driven Integration of Heterogeneous Information Systems

Naumann [15] has developed a system that integrates heterogeneous information systems based on data quality that identify and rank high quality plans, in order to produce results with high quality. The project looks for query plans that are correct and may produce different results while traditional optimization techniques consider plans that all produce same results.

For query processing part, the project process queries by considering the different levels of granularity for each data quality:

- 1) Criteria on the data source, determines the quality of the whole data source, such as timeliness and reputation.
- 2) Criteria on query correspondence assertion (QCA), defines the quality of specific query correspondence assertions, such as the cost of a query.
- 3) Criteria on user query, measures the quality of the answer delivered to a specific user query. The scores for these criteria can only be calculated at query time. An example is completeness.

However, Naumann doesn't specify how to measure these quality criteria at different levels of granularity. The project uses the Data Envelopment Analysis method [16] to rank the data sources. Therefore, user priorities are ignored at this process. Besides, data sources are discarded by subjective criteria such as reputation and understandability.

D. Quality-Driven Query Answering for Integrated Information Systems

Naumann [12] developed a project to integrate data from different data sources based on data quality. The project uses the global relational schema to generate a universal relation to be used for integrating autonomous data sources. Users generates queries by selecting attributes from the universal relation and may specify conditions on the selected attributes. Queries are then transformed to queries against the global relational schema. The project qualifies data sources based on several quality criteria such as objectivity, believability, reputation and others. These criteria are used to generate a quality model for query plans.

The project follows the following steps to calculate the quality of a query plan: Each source gets information quality (IQ) scores for each relevant data quality criteria. The IQ scores are, then, combined to form an IQ-vector. In order for users to determine their preferences, they are required to assign weights to the IQ-vector. Therefore, a weighting vector can be obtained.

A multi-attribute decision-making (MADM) method uses the weighting vector to rank the data sources in the universal relation. An example of MADM method is simple additive weighting (SAW).

After determining the IQ-vector for each data source, the project calculates an IQ-vector for each query plan containing the data sources. Query plans are, then, represented as trees of

joins between the data sources: leaves represent data sources while inner nodes represent the joins between the data sources. By joining nodes from bottom to up, each inner node gets IQ-scores and the overall quality of the plan is the IQ-score of the root of the tree.

IV. QUALITY SYSTEM COMPONENTS

The purpose of adding quality system components to data integration systems is to improve the query answers. This can be achieved by assessing a set of quality criteria over the data sources and storing the measures in a repository to be used later during query planning phase to generate query plans that can produce answers with better quality. These quality system components are: (1) Data quality acquisition and (2) user input. These quality system components are integrated in the mediator-wrapper architecture. See green boxes in "Fig. 4."

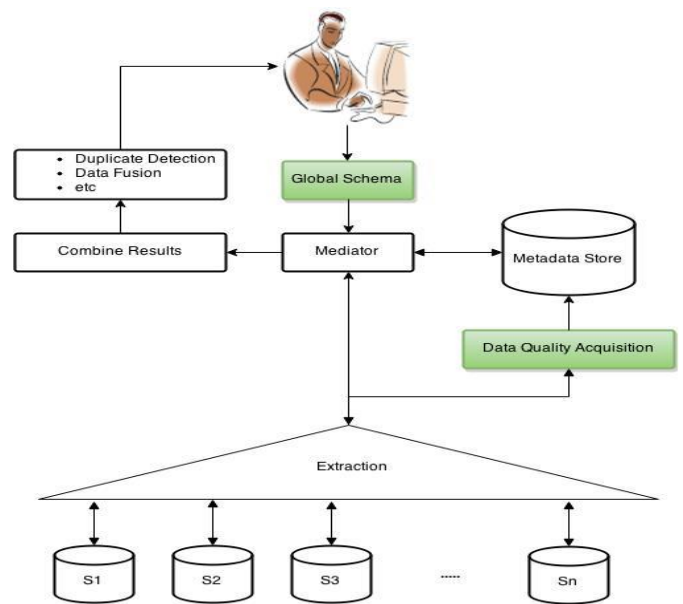


Fig. 4. Data integration system quality system components

In the following sub-sections, we present the structure and the functionality of each component.

A. Data Quality Acquisition

This component is responsible for extracting attributes and relations from the data sources and store them in the metadata store. It is also responsible for executing data quality queries against the data sources, receiving the results and store them in the metadata store. The metadata store used by the data quality acquisition (DQA) consists of the following relations as shown in "Fig. 5":

- 1) Data Sources. Stores information about data sources.
- 2) Tables. Stores information about each relation (table) in each data source. It has an attribute called "detectors" which is used to uniquely identify records in case no primary key exists. This attribute is used during fusion process [17].

- 3) Columns. Stores information about each column in each relation in a data source.
- 4) Global schema columns. Stores information about the global schema attributes.
- 5) Global schema Tables. The global schema columns belong to a global schema table. This makes it easier to add multiple tables with columns.
- 6) Global Schema Mappings. Defines the mapping between the attributes of the global schema and the attributes of the data sources (stored in columns relation) and the mapping functions between the attributes. Such mapping functions are multiplication, division, string concatenation, etc.

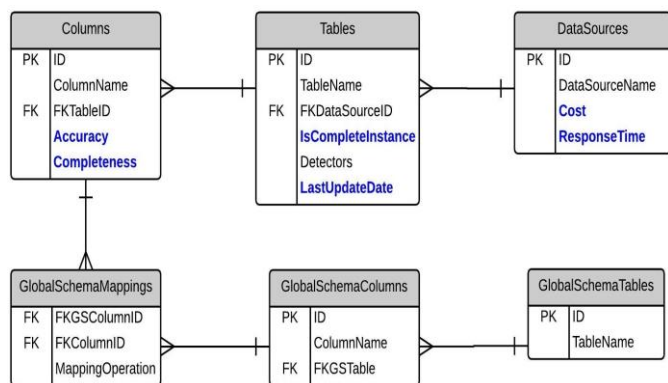


Fig. 5. Metadata structure

Data quality acquisition component can be customized without affecting the data integration system. We can change the queries used by the data quality acquisition anytime. Whenever data quality acquisition executes queries, the quality measures in the metadata store will be updated with the new values.

We have selected a set of data quality dimensions that could affect the data integration process and at the same time could be considered important to the end user to be measured by data quality acquisition. Table I illustrates these dimensions and the granularity for each dimension.

TABLE I. DATA QUALITY DIMENSIONS AND GRANULARITIES LEVELS

DQ Dimension	Measures granularities		
	Data source level	Relation level	Attribute level
Accuracy			✓
Completeness			✓
Cost	✓		
Response time	✓		
Timeliness		✓	

In the following sub-sections, we describe how we measure each dimension presented in table I. These quality measures may enhance the quality of the data fusion process. Data quality dimensions chosen are highlighted in blue in “Fig. 5”.

1) Accuracy

Tomas C. Redman [10] present the data accuracy measurement framework (“Fig. 6”) for understanding the various measurement techniques based on choices made regarding four factors: where to measure the data, which part of the data will be measured, how to measure the data and the granularity of the measures.

To apply Redman’s data accuracy measurement framework in our case, we will select from the choices for each of the four factors.

- Where measurements are taken: Since we have a set of data sources given, we will measure accuracy from those data sources. (i.e. from database).
- What attributes to include: To save processing time, we measure accuracy on the data sources’ attributes that correspond to global schema’s attributes.
- The measurement device: Since a reference to a real world relation is almost always costly and time consuming, we will compare the value of each attribute to its domain of allowed values. Complaints and domain experts’ feedback are also used to identify erred data and a correction for them which help improve accuracy measure.
- The scale on which results are reported: Attribute level.

$$\text{Attribute Accuracy} = \frac{\text{Number of fields judged correctly}}{\text{Number of fields tested}} \quad (1)$$

2) Completeness

The Literature classifies completeness into three types: column completeness, schema completeness, and population completeness [18]. At the most abstract level, schema completeness refers to the degree to which all required information are present in a particular data set. At the data level, column completeness can be defined as the measure of the missing values for a column in a table. Each of the three types can be measured by dividing the number of incomplete items by the total number of items and subtracting from 1 [18].

$$\text{Schema/Attribute completeness} = 1 - \frac{\text{Number of incomplete items}}{\text{Total number of items}} \quad (2)$$

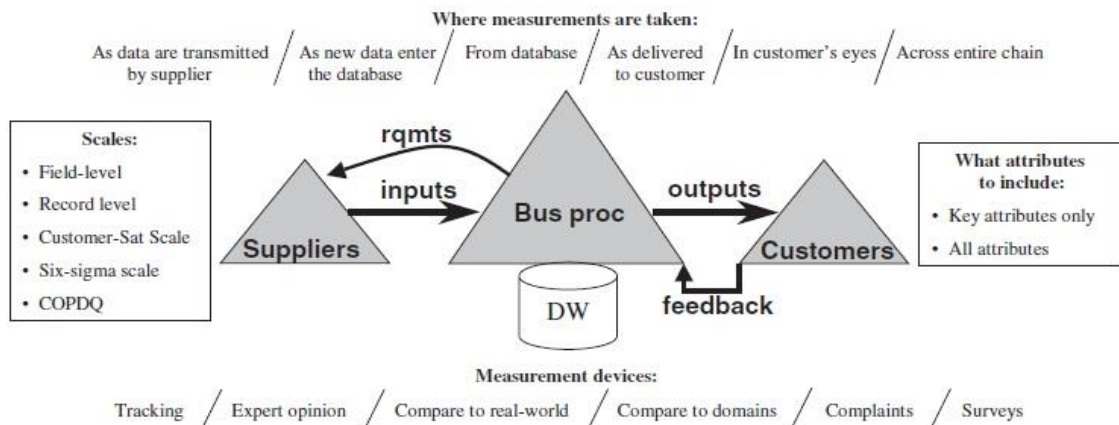


Fig. 6. The data accuracy measurement framework

The range for completeness is 0 - 1, where 0 represents the lowest score and 1 represents the high score.

In relational databases research area, completeness is often related to the meaning of “null” values. A value can be missing because it doesn’t exist or because it exists, but is not known or because its exist and not identified. We apply completeness on non-primary key attributes and applicable attributes. We add a custom data quality criteria called “Complete instance relation” that can be measured at schema level. A relation is marked as complete instance if its cardinality is complete. (i.e. all the tuples are represented in the relation). This information will be given directly to the data integration system by end user through an input form.

3) Cost

It is the price for accessing specific data source. The user has to pay the money for accessing a commercial data source. We added this criteria due to the growing importance of commercial data source providers. The subscription of the user and the cost of that subscription are determined by the data source owner. We assume that the user is charged on pay-by-query basis. The cost per query is measured in US dollar.

4) Response Time

We measure the response time of a data source by using calibration techniques [19]. We send a bunch of queries to the data sources to judge their average response time for different types of queries at different times of day. The result will be stored at the metadata store to be used later during query planning phase. The response time may be high if a source is always busy or doesn’t have the resources needed to answer the query. In this study, we assume that all the data sources have the capabilities to answer all queries so that the problem of source capabilities is resolved. The response time depends on several factors such as network traffic, servers workload, technical equipment such as: the hardware of the data source and database management system used by the data source.

5) Timeliness

Timeliness measure depends on the data integration system: some prefer seconds while others prefer days. To determine the

timeliness, we rely on update information provided by the data source. Timeliness measurement depends on at what granularity the data source updates its data. We assume at the relation level and the data at the data sources are not archived.

B. User Input

To give users the option to specify constraints on the retrieved result, we have used the proposal of Gertz and Schmitt [14] where quality constraints can be expressed by using data quality dimensions. Thus forming a threshold of acceptance. We have added two options to the SQL dialect. The first one is cost which is the amount of money a user can pay and the second option called fusion that can be set to true or false and is used to give the user the option to fuse data from all possible data sources.

A query Q with quality constraint expressed on the mediated schema expressed in an extended SQL syntax:

```
Select A1,.....,Ak
from G
where < selection condition >
with < data quality goal >
fusion < true | false >
Cost < x$ >
Where A1.A2,...., Ai are global attributes of G
```

Selection condition: conditions used to filter the data.

Data quality goal: quality dimensions defined on the selected attribute Ai and gets a value according to table II.

TABLE II. DATA QUALITY DIMENSIONS LEVELS

Level	Start threshold
High	70
Meduim	50
Low	0

The values in table II are used as a threshold of acceptance. Ex: if we have a quality constraint A1.completeness is Medium, that’s mean that the user wants the completeness of A1 in the answer to be 50% or higher.

The values in table II are tunable. The system administrator can change the value each level anytime.

Fusion: When set to true, this means that the user wants to fuse data from all possible data sources. When set to false, the mediator selects only one alternative that has the minimum number of data sources.

Cost: the amount in US dollar the user can pay.

V. QUALITY DRIVEN QUERY PROCESSING ALGORITHM

The requested data usually located on more than one data source. Every combination of data sources that meet the user's requirements (attributes and quality criteria) is an alternative. If a single data source can meet all user's requirement, this is an alternative. Given a query Q against the mediated schema asking for A1,.....,An attributes with or without quality requirements, We developed a quality-driven query algorithm ("Fig 7") to determine which combinations of sources can answer the query. The algorithm works as follows:

- 1) Since each global schema attribute is assigned a unique ID, the mediator obtains a metadata representation by joining the tables in the metadata store and using the IDs of the selected attributes to retrieve the metadata related to the selected attributes only.
- 2) After determining the data sources that can answer the query, we examine the quality criteria required by the user (if exist) against the quality of these data sources. We discard the data sources that don't meet user's cost criteria. Ex: If the user can pay x\$, then all data sources whose cost greater than x, will be discarded.
- 3) The mediator extracts the complete instance relations from the remaining data sources and examines the quality and the attributes provided by these relations against the quality and the attributes required in the user's query. This process is as following:

```

Input: metadata representation
Output: list of alternatives
// means comment

1: List_of_alternatives = {}
2: List_of_warningMessages = {}
3: temporary_alternatives = {}
4: missing_attributes(key, value) = {} // Key represents the missing or disqualified attribute while
// value represents the data sources that provide that attribute.

5: for i = |complete_instance_relations| to 1 do
6:   Set R = complete_instance_relations[i];
7:   if R contains all required attributes && R meet all DQ criteria then
8:     List_of_alternatives.push(data source of R);
9:   else
10:    missing_attributes = {};
11:    Set A = the missing or disqualified attributes from R;
12:    for m = 1 to |A| do //add the missing or disqualified attributes to the collection
13:      missing_attributes.push(A[m]);
14:    end for
15:    for m = |complete_instance_relations| to 1 do
16:      Set k = complete_instance_relations[m];
17:      for z = |missing_attributes| to 1 do
18:        if k contains A[z] && k meet the DQ criteria required for A[z] (if exist) then
19:          missing_attributes[A[z]] += the data source of k; //Add the data source of K to A[z]
20:        end if
21:      end for
22:    end for
23:    //Sort missing_attributes in descending order based so that the attribute that has the highest number of data sources
24:    //becomes first
25:    temporary_alternatives = {};
26:    if fusion = true then
27:      Set D = missing_attributes[1]; // set D = the data sources for the first element
28:      for f = |D| to 1 do
29:        temporary_alternatives.push(data source of R with D[f]);
30:      end for
31:      for f = 2 to |missing_attributes| do
32:        for w = 1 to |temporary_alternatives| do
33:          Set D = missing_attributes[f]; // D = data sources that provide missing_attributes[f]
34:          temporary_alternatives[w] += D; // update each alternative with the data sources that provide
35:          //missing_attributes[f] so that the alternative meets the user's needs.
36:        end for
37:      end for
38:      for w = 1 to |temporary_alternatives| do
39:        List_of_alternatives.push(temporary_alternatives[w]);
40:      end for
41:    else // Generate alternative for each missing attribute
42:      temporary_alternatives.push(data source of R);
43:      for f = 1 to |missing_attributes| do
44:        Set D = missing_attributes[f]; // D = data sources that provide missing_attributes[f]
45:        if |D| > 1 then
46:          Set E = DataEnvelopmentAnalysis(D); // Apply DEA on the data sources and set the
47:          //efficient data sources to E
48:          for w = 1 to |E| do
49:            temporary_alternatives[1] += E[w];
50:          end for
51:        else
52:          temporary_alternatives[1] += D;
53:        end if
54:      end for
55:    end if
56:    List_of_alternatives.push(temporary_alternatives[1]);
57:  end if
58: end for
59: if |List_of_alternatives| == 0 then //No data source found that can match the required DQ criteria
60:   List_of_warningMessages.push("Sorry, no data sources found match the DQ criteria required for attribute x");
61:   List_of_alternatives.push(data source of R); // we must provide answer. Even partial answer
62: end if

```

```

59: if |complete_instance_relations| == 0 then           //when no instance relations found
60:   Set R = incomplete_instance_relations[1];
61:   List_of_alternatives.push(data source of R);
62:   for i = 1 to |incomplete_instance_relations| do
63:     Set R = incomplete_instance_relations[i];
64:     List_of_alternatives[1] += data source of R
65:   end for
66: end if
67: if fusion = true then
68:   Merge the data sources in all alternatives and query each data source only once
69: else
70:   Select the alternative that has the minimum number of data sources and query the data
71:   sources in it.
72: end if
73: set DS = number of data sources queried;
74: if |DS| > 1 then
75:   //Apply duplicate detection algorithm
76:   //Apply data fusion algorithm
77: end if
78: Display the result to the user
    
```

Fig. 7. Quality driven query algorithm

- a) If a relation R in data source S doesn't provide at least one attribute then the mediator examines other data sources for that missing attribute.
 - b) If a relation R in data source S doesn't meet the quality criteria required for at least one attribute then the mediator examines other data sources for that disqualified attribute.
 - c) If a relation R in data source S provides all required attributes and meets all the quality criteria required in Q, then no other data sources are needed and the mediator adds the data source of R to the list of alternatives.
 - d) The mediator applies the above three steps for the remaining complete instance relations.
- 4) If no complete instance relation is found in step 3, then all data sources will be used to form an alternative.
- 5) The following steps illustrate how the mediator finds other data sources that provide the missing attributes or the disqualified attributes:
- a) The mediator checks the metadata representation for data sources that provide the required attributes.
 - b) If no relations found in step 5-a, that means that one of the required quality criteria can't be factory. So, a warning message will be displayed to the user regarding that quality criteria unless an alternative is found. However, the mediator must provide an answer. So, we add the data source of R (R is defined in step 3) to the list of alternatives.
 - c) If relations are found in step 5-a, then the mediator checks if fusion option in Q is set to true or false.
 - If fusion is set to true, the mediator sorts the missing attributes in a descending order based on the number of data sources found for each attribute. The reason for that is to make sure that for each data source S1, S2, ..., Sn that provide the first missing attribute, the mediator generates alternatives that consists of the data source of R (R is defined in step 3) and Si. Then the mediator iterates on the remaining attributes and updates the alternatives with data sources that provide each missing attribute. Finally, the alternatives are then added to the list of alternatives.
 - If fusion is set to false, the mediator generates only one alternative for all required attributes. To find that alternative, for each missing attribute A1, A2, ..., Ai, we apply Data Envelopment Analysis (DEA) method on the data sources that provide Ai and any efficient source will be added with the data source of R to the list of alternatives. If no efficient data sources are found, the highest inefficient data source will be added with the data source of R to the list of alternatives. The quality measures, which are already stored in the metadata store, will be used in DEA. These quality measures are: Completeness, Accuracy, Cost and Response time.
- 6) After generating all alternatives, the mediator checks if fusion option in Q is set to true or false.
- If fusion is set to true, the mediator merges the data sources in all alternatives and query each data source only once.
 - If fusion is set to false, the mediator selects the alternative that has the minimum number of data sources and query the data sources in it.
 - If the number of queried data sources is greater than one, duplicate detection and data fusion algorithms will be run on the result respectively.
 - The result is then displayed to the user.

VI. EVALUATION AND VALIDATION

In this section, we validate that our approach really reduces the number of data sources needed to answer a given query.

Given the Student schema in data sources S1, S2, S3, S4, S5, S6 and S7. The mediated schema is shown below

G: Student (FirstName, LastName, Gender, Birthdate, Mail, Nationality, Address, Phone)

S1.Student (StudentID, FirstName, LastName, Gender, Birthdate, Address)

S2.Students (SID, Name, Sex, Birthdate, E-mail, Nationality)

S3.Student (ID, FullName, Email_Address, Nationality, Phone)

S4.Student (Id, FName, LName, Gender, Mail, Nationality, Address, Phone)

S5.Student (Student_id, Student_Name, Gender, Nationality, Birthdate)

S6.Student (Student_ID, Name, Gender, Mail, Phone)

S7.Student (S_ID, FName, LName, Sex, Address, Birthdate)

The data sources measures in the metadata store are shown in table III.

TABLE III. DATA SOURCES MEASURES IN THE METADATA STORE

Data Source ID	Properties		
	Data Source Name	Response Time	Cost
1	S1	92 sec	10\$
2	S2	160 sec	5\$
3	S3	130 sec	3\$
4	S4	280 sec	0\$
5	S5	500 sec	0\$
6	S6	350 sec	7\$
7	S7	300 sec	0\$

Now, Consider the following query Q1:

```
Select FirstName, LastName, Gender, Birthdate
From G
Where Gender ="Male"
With Birthdate.completeness is high
fusion = true
Cost = 0
```

The interpretation of the above query is that the user wants First Name, Last Name, Gender and Birthdate of all male students where completeness of birthdate is high (i.e. completeness measure starts from 70 as indicated in table II) and obtain the result from the free data sources only.

The first step to process the above query is by obtaining metadata representation by joining the tables in the metadata store (see "Fig. 5") and filtering the rows by the IDs of the selected attributes to retrieve the metadata related to the selected attributes only.

Second, we discard the data sources that don't meet the cost criteria specified in Q. Therefore, S1, S2, S3 and S6 are discarded. This yields the metadata representation shown in table IV.

TABLE IV. METADATA REPRESENTATION

Properties					
Column Name	Table	Accuracy	Completeness	Table Name	Complete instance
FName	S4	99	100	Student	Yes
LName	S4	97	100	Student	Yes
Gender	S4	100	100	Student	Yes
Student_Name	S5	88	95	Student	Yes
Gender	S5	100	100	Student	Yes
Birthdate	S5	80	84	Student	Yes
FName	S7	100	100	Student	Yes
LName	S7	100	100	Student	Yes
Sex	S7	100	100	Student	Yes
Birthdate	S7	80	90	Student	Yes

Third, the mediator extracts the complete instance relations from the remaining data sources. From table IV, complete instance relations are S4, S5 and S7. The mediator starts with S4 and finds that S4 does provide all required attributes except Birthdate, so the mediator will look for other data sources that could provide attribute "Birthdate". After scanning the metadata representation (table IV), the mediator finds S5 and S7. The mediator then checks the fusion option in Q. Since fusion option is set to true, the mediator generates an alternative for each data source. So, the list of alternatives = {{S4, S5}, {S4, S7}}.

Next, the mediator examines S5 and finds that S5 does provide all required attributes (S5.Student.Student_Name contains FirstName concatenated with LastName). So no other data sources are needed. Therefore, S5 itself is an alternative and the mediator will add it to the list of alternatives. Therefore, the list of alternatives = {{S4, S5}, {S4, S7}, {S5}}.

Next, the mediator examines S7 and found that S7 does provide the required attributes. So, no other data sources are needed. Therefore, S7 itself is an alternative and the mediator adds it to the list of alternatives. Therefore, the list of alternatives = {{S4, S5}, {S4, S7}, {S5}, {S7}}.

Now, given a set of alternatives, the mediator determines which alternatives to choose as follows:

- The mediator checks again fusion option in Q. Since the fusion is set to true, the mediator merges the data sources in all alternatives and query each data source only once. Therefore, the final query plan = {S4, S5, S7}.
- After retrieving the result from each data source, the mediator unions the results and applies duplicate detection algorithm to find the tuples that refer to the same real world entity.

- After determining the duplicate records, a data fusion algorithm is needed to fuse attribute's values that refer to the same real world entity (resolve inconsistency at value level).
- Any further processing on the result can be applied.
- Finally, the result is displayed to the user.

If we assume the complete instance property for S4 is “No”, we will find that the final list of alternatives {{S5}, {S7}} and the final query plan as {S5, S7}. S4 is omitted because it doesn't provide a complete result. This reduces the time needed to answer a query by avoiding access to S4.

If we modify fusion option and set it to false and repeat the above steps, we will find that the metadata representation is still the same (table IV). S4 does provide all required attributes except Birthdate. The mediator selects S5 and S7 to provide attribute Birthdate. Since fusion is set to false, the mediator tries to choose the best source between S5 and S7 to retrieve attribute birthdate from. The mediator achieves this by applying DEA on S5 and S7. The quality scores in table V are used in DEA.

TABLE V. QUALITY SCORES

Data Source	Quality Criteria			
	Accuracy	Completeness	Response Time	Cost
S5	80	84	500 sec	0\$
S7	80	90	300 sec	0\$

The computed efficiency for S5 is 0.7939 while the computed efficiency for S7 equals 1. Therefore, the mediator adds S7 along with S4 as an alternative to the list of alternatives. The list of alternatives = {{S4, S7}}.

Next, the mediator examines S5 and S7 respectively and finds both provide all required attributes. Therefore, S5 and S7 themselves are alternatives and the mediator adds them to the list of alternatives. Therefore, the list of alternatives = {{S4, S7}, {S5}, {S7}}.

Now, given a set of alternatives, the mediator determines which alternatives to choose as the following:

- Since the fusion is set to false, the mediator chooses the alternative that has the minimum number of data sources and query the data sources in it. In this case, the mediator can choose either S5 or S7. Therefore, the final query plan = {S7}.
- Since the number of data sources queried equals one, neither duplicate detection algorithms nor data fusion algorithms are needed unless the data source allows duplicate.
- Any further processing on the result can be applied.
- Finally, the result is displayed to the user.

VII. CONCLUSION AND FUTURE WORK

Data integration systems may produce query results that not only suffer the lack of quality but also take a long time to arrive.

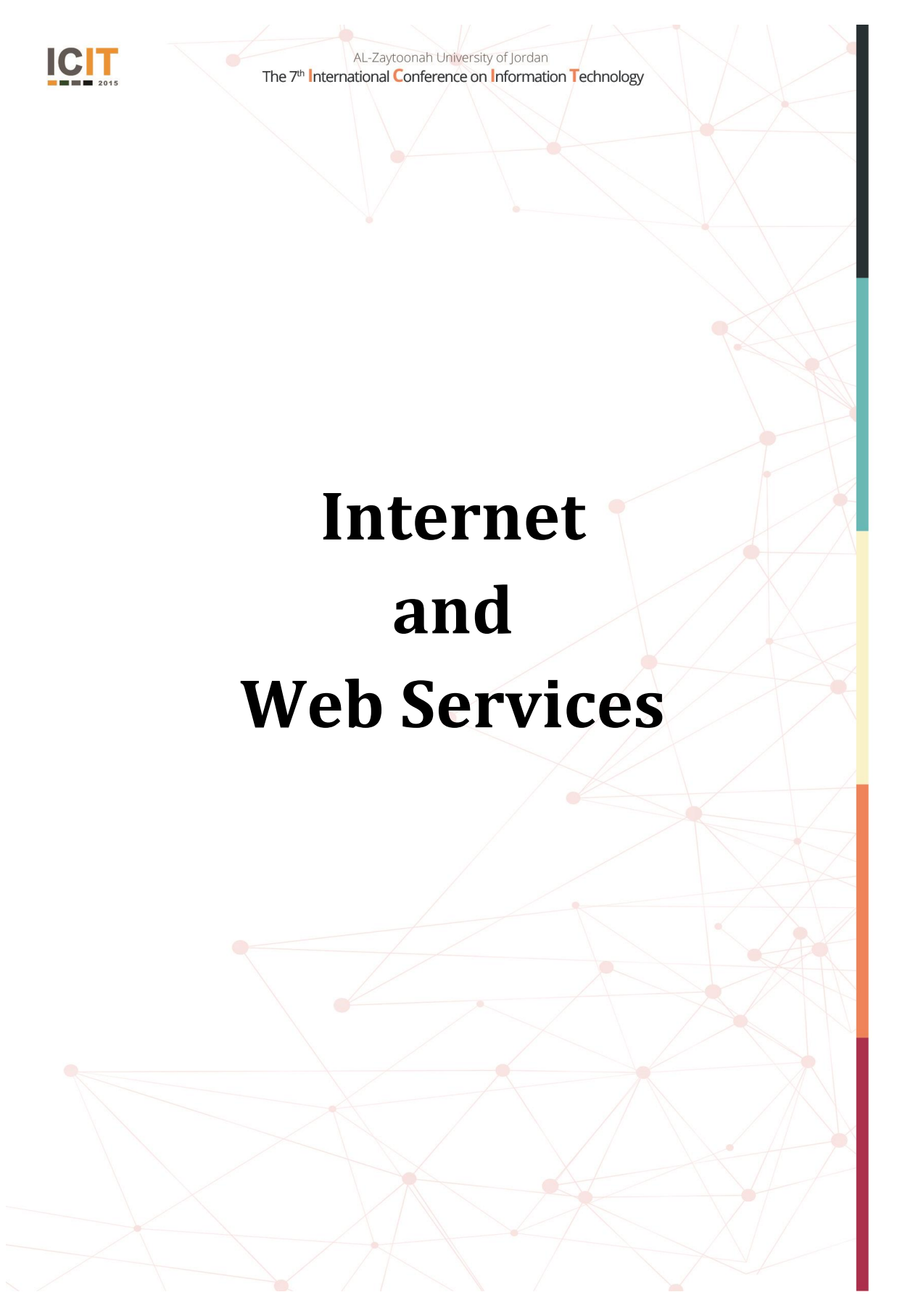
The results can be incomplete, inaccurate or outdated and so on. In this paper, we have pointed out the importance of data quality in integrating autonomous data sources. The main contribution of this paper is an efficient method aimed at selecting a few possible data sources to provide more quality oriented result to the user. We added quality system components to integrate data quality dimensions in a data integration environment for structured data sources only. With the help of these criteria, we developed a quality driven query execution algorithm to generate high quality plan that meets user's requirements. Further research will extend the approach to be applied on different types of data sources such as semi-structured and unstructured data sources.

REFERENCES

- [1] M. Lenzerini, "Data integration: a theoretical perspective," in Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database system, Madison, Wisconsin, June 03-05, 2002.
- [2] A. Y. Halevy, "Answering queries using views: A survey," The VLDB Journal — The International Journal on Very Large Data Bases, vol. 10, no. 4, pp. 270-294, December 2001.
- [3] A. Doan, A. Halevy and Z. Ives, Principles of Data Integration, San Francisco, CA: Morgan Kaufmann Publishers Inc., 2012.
- [4] G. Wiederhold, "Mediators in the Architecture of Future Information Systems," IEEE Computer, vol. 25, no. 3, pp. 38-49, March 1992.
- [5] Y. Wand and R. Y. Wang, "Anchoring data quality dimensions in ontological foundations," Communications of the ACM, vol. 39, no. 11, pp. 86-95, November 1996.
- [6] M. Ge and M. Helfert, "A Review of Information Quality Research - Develop a Research Agenda," in Proceedings of the 12th International Conference on Information Quality (ICIQ 07), MIT, Massachusetts, USA, November 9-11, 2007.
- [7] J. M. Juran, The Quality Control Handbook, 3rd ed., New York: McGraw-Hill, 1974.
- [8] G. Kumar Tayi and D. P. Ballou, Examining data quality, Communications of the ACM., v.41 n.2, p.54-57, Feb. 1998.
- [9] R. Y. Wang and D. M. Strong, "Beyond accuracy: what data quality means to data consumers," Journal of Management Information Systems, vol. 12, no. 4, pp. 5-33, March 1996.
- [10] T. C. Redman, "Measuring Data Accuracy A Framework and Review," in Information Quality, R. Y. Wang, E. M. Pierce, S. E. Madnick and C. W. Fisher, Eds., Armonk, NY, M.E. Sharpe, 2005, pp. 21-36.
- [11] C. Batini and M. Scannapieco, Data Quality: Concepts, Methodologies and Techniques (Data-Centric Systems and Applications), Secaucus, NJ: Springer-Verlag New York, Inc, 2006.
- [12] F. Naumann, Quality-Driven Query Answering for Integrated Information Systems, Springer-Verlag Berlin, Heidelberg, 2002.
- [13] M. Scannapieco, A. Virgillito, C. Marchetti, M. Mecella and R. Baldoni, "The daquincis architecture: a platform for exchanging and improving data quality in cooperative information systems," Information Systems, vol. 29, no. 7, pp. 551 - 582, October 2004.
- [14] M. Gertz and I. Schmitt, "Data integration techniques based on data quality aspects," in 3rd National Workshop on Federal Databases, Magdeburg, Germany, 1998.
- [15] F. Naumann, U. Leser and J. C. Freytag, "Quality-driven integration of heterogenous information systems," in 25th proceeding of the International Conference on Very Large Databases (VLDB) , p.447-458, Edinburgh, Scotland, September 07-10, 1999.

- [16] A. Charnes, W. W. Cooper and L. Rhodes, "Measuring the efficiency of decision making units," *European Journal of Operational Research*, vol. 2, no. 6, pp. 429-444, November 1978.
- [17] A. Z. El Qutaany, A. H. El Bastawissy and O. Hegazy, "A Technique for Mutual Inconsistencies Detection and Resolution in Virtual Data Integration Environment," in *Informatics and Systems (INFOS)*, 2010 The 7th International Conference on, 2010.
- [18] L. L. Pipino, Y. W. Lee and R. Y. Wang, "Data quality assessment," *Communications of the ACM*, vol. 45, no. 4, pp. 211-218, April 2002.
- [19] M. Spiliopoulou, "A calibration mechanism identifying the optimization technique of a multidatabase participant," in *Proc. of the Conf. on Parallel and Distributed Computing Systems (PDCS)*, Dijon, France, September 1996.

Internet and Web Services



Privacy and Protection in Electronic Transaction: A Review of The E-Commerce Security Protocols

Taroub Ahmed Mustafa Sa'ed
Faculty of Technology and Applied Science
Al-Quds Open University
Palestine
tessa@qou.edu

Abstract—E-commerce applications are becoming popular day by day as they are working like a virtual shop. Writing good E-commerce application is tedious task and complex also. The applications if made complex are very difficult to maintain. Usability is a very basic concept in the E-commerce application. User has to get the information at one click and with proper feedback. As these are web based applications, efficiency matters a lot for this application. As transaction in e-commerce faces the problems such as database exploits, log data mining and sniffing attacks which can be resolved by using different security measure. Hence, security is important in e-commerce application. In today's electronic world of business, trust is the center component between the consumer and the internet Merchant. Researchers found trust very important, especially, in the relationships between consumers and e-vendors. Based on the analysis of the basic concepts, the security infrastructure and payment system of electronic commerce and the thorough and comprehensive research on the security technology, authentication and transaction process, this paper points out some aspects of excellence and deficiencies in security protocols beginning with iKP, SSL, SET, 3d-Secure and finally other modified models of 3d-secure protocols.

Keywords: *Electronic Commerce, SET, Sniffing Attack, Log Data Mining, DBMS exploit, DES, RSA, 3d-Secure, SSL.*

I. INTRODUCTION

The internet is a network of networks. Connecting a business to the internet implies a global reach. In other words, a company can reach anyone who has an access to the internet such as customers, suppliers, on-line banks, mediators, etc. At the same time, the company can be reached by anyone. So, the internet creates vast opportunities for businesses with some threats. For example, anybody from anywhere on the internet (an intruder or a hacker) can illegally enter a company computer resource and messes the computer resource from a remote site. In addition, it is not difficult task to tap a message in the middle of the net and steal or change its content, which is definitely a crime. While the internet is dramatically changing the way business is conducted, security issues are of deeper concern than ever before. The internet is basically an insecure communication medium. Hawker [6] states that the only assumption which can safely be made when considering the internet as a communication medium is that it offers no security whatsoever. Most people are skeptical about the security of the internet. People are happy using the World Wide Web for browsing, finding, reading or downloading information from the internet. However, when considering sending a credit card number over the internet, they are reluctant to do it, even if they are told that the transfer is secured. This is because many media expose bad news about the internet

security, although security technology for the internet exists and good enough for protecting transactions via the internet. The core activities of e-commerce are business transactions between two parties or possibly mediated by a third party. In fact, the practice conducted by company before the term e-commerce appears is Electronic Data Interchange (EDI), which is basically electronic transaction via computer networks. The major concern of electronic transactions is how to protect transactions from eavesdroppers (which can steal and modify the information in the transactions) and how to make sure those transactions are authenticated. This paper begins by outlining basic concepts in section 2, while section 3 deals with protocols in Electronic Transaction like SSL, SET and 3d-secure protocols. Also other modified models of 3d-Secure will be discussed. Finally, the conclusion, discussion and references.

II. BASIC CONCEPTS

A. Security Issues in E-Commerce Application

E-commerce was established in 1991; it is selling and buying of products and services by business and consumers via computer network such as internet. From the year 1991 till today, life has faced drastic changes due to technological advancements, but simultaneously they have made our lives more complex. Moreover, E-Commerce became a need for development and modifying to protect the data and information from any attack, from here the definition of security in EC has emerged[4].

Electronic commerce lets companies integrate internal and external business processes through information and communication technologies. Companies conduct these business processes over intranets, extranets, and the internet. E-commerce lets businesses reduce costs, attain greater market reach, and develop closer partner relationships. However, using the internet as the underlying backbone network has led to new risks and concerns. Often, industry analysts considered trust and security as the main hurdles in growing e-commerce. A number of factors have hampered the growth of e-commerce in developing countries. Yet, the main perceived obstacle to increased internet usage is very similar in companies from both developed and developing countries. Firms already using the internet consider the lack of network security to be the primary problem, followed by slow and unstable connections. This litany of evolutionary phases masks a number of growing technical challenges, which could be addressed as the following[17][16]:

- security and authentication.
- content management and publication.
- reliable systems, messaging, and data.
- complex interactions and transactions.
- business model implementation and business process enactment.
- distributed processing and distributed data.

Clearly, the online transaction requires consumers to disclose a large amount of sensitive personal information to the vendor, placing themselves at significant risk. Understanding (indeed, even precisely defining) consumer trust is essential for the continuing development of e-commerce.[12].

There are different types of security issues in any e-commerce application which needs to be addressed as the following[1][11]:

- 1) Malicious Code such as Viruses.
- 2) Unwanted Programs: These are installed without the user's informed consent. It has three types:
 - Browser parasites: It can monitor and change settings of a user's browser.
 - Adware: It calls for unwanted pop-up ads.
 - Spyware: It can be used to obtain information, such as a user's keystrokes, e-mail, IMs, etc.
- 3) Phishing and Identity Theft: which means that any deceptive, online attempt by a third party to obtain confidential information for financial gain – Most popular type: e-mail scam letter – It is one of fastest growing forms of e-commerce crime
- 4) Hacking and Cyber vandalism.
- 5) Credit Card Fraud.
- 6) Spoofing (Pharming) and Spam (Junk) Web Sites.
- 7) Denial of service (DoS) attack and Distributed denial of service (DDoS) attack.
- 8) Other Security Threats: like:
 - Sniffing: Type of eavesdropping program that monitors information travelling over a network; enables hackers to steal proprietary information from anywhere on a network

- Insider jobs: Single largest financial threat.
- Poorly designed server and client software: Increase in complexity of software programs has contributed to increase in vulnerabilities that hackers can exploit.

B. Mechanisms and Technologies to Build Trust

Trust is especially an important factor under conditions of uncertainty and risk. The importance of trust is elevated in e-commerce because of the high degree of uncertainty and risk present in most on line transactions. In today's electronic world of business, trust is the center component between the consumer and the internet Merchant. Researchers found trust very important, especially, in the relationships between consumers and e-vendors[17].

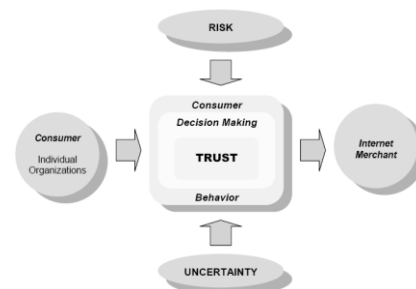


Fig 1: A Relationship between Consumer and internet Merchant[17].

There is a strong relation between consumer trust and security aspects that govern the whole transaction processes in an e-commerce website.

As a new form of commercial activity, e-commerce involves more uncertainty and risks than traditional commerce because they are less well known to consumers. Factors that affecting trust in e-commerce for consumers include security risks, privacy issue and lack of reliability e-commerce processes in general. A consumer cannot monitor the safety and security of sending sensitive personal and financial information. Online business organizations should search for high-tech security mechanism to protect itself from intrusion and also protect it's customer from being indirectly invaded. There are two lines of defense for e-commerce which are technology solutions and policy solutions, here we state some of adapted solutions: [1][11][3].

1) Encryption Approach: Encryption is the process of transforming plain text or data into cipher text that cannot be read by anyone other than the sender and the receiver. It's purpose is:

- (a) to secure stored information.
- (b) to secure information transmission..

There are several types of encryption that differs in the context of it's functionalities. In Symmetric Key Encryption, both the sender and the receiver use the same key to encrypt and decrypt messages while Public Key Encryption used two mathematically related digital keys which are public key and private key[11][3].

2) Secure Socket Layer: The most common form of securing channels is through the Secure Sockets Layer (SSL) of TCP/IP. The SSL protocol provides data encryption, server

authentication, optional client authentication, and message integrity for TCP/IP connections. Secure Socket Layer (SSL) is a security protocol, first developed by Netscape Communications Corporation and now taken over by the transport layer security working groups. The design goal of the protocol is to prevent eavesdropping, tampering or message forgery when a data is transported over the internet between two communicating applications[3].

3) Secure Hypertext Transfer Protocol (S-HTTP): S-HTTP is a secure message-oriented communications protocol designed for use in conjunction with HTTP. It is designed to coexist with HTTP and to be easily integrated with HTTP applications. Whereas SSL is designed to establish a secure connection between two computers, S-HTTP is designed to send individual messages securely. Using S-HTTP, any message may be signed, authenticated, encrypted or any combination of these. Generally, S-HTTP attempts to make HTTP more secure[3].

4) Trust Seals Programs: A number of Trustmark seals have been developed to provide assurance about Web business practices and policies through the Web interface. One example is TRUSTe, which audit a site's stated privacy policies and allows sites to display the TRUSTe seal if privacy policies and disclosure meet specific standards. Cheskin and Sapient [5], found that where trust mark seals were recognized, they increase consumer perceptions of a site's trustworthiness.

Seal programs such as TRUSTe, BBBOnLine, MultiCheck and WebTrust allow licensees who abide by posted privacy policies and/or allow compliance monitoring to display means for addressing consumer privacy concerns.

5) Digital Signature: Digital signature means a digital method executed by a party with the intent to authenticate a record, which is a unique to the person using it and is capable of verification. It is linked to the data in such a manner that if the data is changed, the electronic signature is invalidated. A digital signature is normally a hash of the message which is encrypted with the owner's private key[11][3].

6) Secure Electronic Transaction (SET):

A SET specification for credit/payment card transactions is required for the safety of all involved in e-commerce. It is designed to meet three main objectives. First, it will enable payment security for all involved, authenticate card holders and merchants, provide confidentiality for payment data and define protocols and potential electronic security service providers. It will also enable interoperability among applications developed by various vendors and among different operating systems and platform[11][3].

7) Privacy Policy Statements: A privacy policy statement is a contractual commitment to consumers outlining how their personal information will be treated. The evidence suggests that posting a self-reported guarantee of compliance with e-commerce standards is an effective means of increasing consumer trust. Privacy policy statements appear to be most beneficial to the web merchants that have the greatest need to

increase consumer trust . Privacy is the willingness of consumers to share information over the internet that allows purchases to be conducted.

8) Digital Certificate: A digital certificate is a digital document issued by a trusted third party institution known as a certification authority that contains the name of the subject or company, the subject's public key, a digital certificate serial number, an expiration date, an issuance date, the digital signature of the certification authority and other identifying information. The Certification Authority (CA) is a trusted third party that hands out certificates and publishes identities and public keys in a directory. The certificate is signed with the private key of the Certification Authority; therefore, its authenticity can be confirmed by using the known public key of the CA[11][3].

C. Encryption Technology

Encryption is the key security schemes adopted for electronic payment systems, which is used in protocols like SSL and SET. It is a very old technology for keeping messages secret from unauthorized access. One of the oldest methods of encryption was developed by Spartan generals around the fifth century of BC [6]. The basic idea of encryption is only an authorized person can reveal information from an encrypted message by using a key.

Encryption algorithms can be sited to two types, namely symmetric cryptography (single key cryptography) and asymmetric cryptography or public key (two keys) cryptography. A well-known symmetric cryptography is DES (Data Encryption Standard) developed by IBM for the US government. A well-known public key cryptography is RSA cryptosystem.

Data Encryption Standard (DES) is the most widely used symmetric cryptography. DES was adopted by NIST (National Institute of Standards and Technology) in 1977 to provide an encryption algorithm to be used in protecting federal unclassified information from unauthorized disclosure or undetected modification during transmission or while in storage [10].

The DES algorithm uses a 56-bit key to encrypt plaintext to ciphertext or to decrypt ciphertext to plaintext. The DES consists of 16 "rounds" of operations that mix the data and key together in a prescribed manner using the fundamental operations of permutation and substitution. The goal is to completely scramble the data and key so that every bit of the ciphertext depends on every bit of the data plus every bit of the key.

The RSA Cryptosystem is an asymmetric cryptosystem developed by the trio: Ronald Rivest, Adi Shamir and Leonard Adleman [13]. The RSA cryptosystem is based on the principle that if two large prime numbers are multiplied, the resulting number is hard to factor back to its original numbers. In the RSA cryptosystem the two numbers are keys, namely private and public keys. A private key must be kept secret, while a public key can be revealed to anyone.

Obviously, the RSA cryptosystem is more complex and harder to manage than DES since it involves two keys. However, an inherent benefit will be revealed shortly.

In the RSA cryptosystem, a sender may encrypt a message using his/her private or public key. Let A and B be two parties that use the RSA cryptosystem and KPA, KTA, be the public key and the private key for A, KPb, KTB be the public key, the private key for B respectively. Assume that B knows KPA and A knows KPb. There are two possible scenarios:

1. A sends a message to B. Before sending the message, A encrypts the message using KPb. Since A uses KPb to encrypt the message then only B can decrypt the message using KTB. This is called the encryption path of the RSA cryptosystem.

2. A sends a message to B. Before sending the message, A encrypts the message using, KTA. Next, B decrypts the message using KPA. If B can decrypt the message using KPA, then the message must come from A. This is called the authentication path, which can be used as a digital signature (the message is digitally signed by A). Note that A cannot deny (non-repudiation principle) that he/she has signed the message since the message can only be decrypted using A's public key (KPA) [13].

III. SSL, SET AND 3D-SECURE: PROTOCOLS OF ELECTRONIC COMMERCE

Electronic commerce, as exemplified by the popularity of the internet, is going to have an enormous impact on the financial services industry. No financial institution will be left unaffected by the explosion of electronic commerce. Most people are skeptical about the security of the internet. People are happy using the World Wide Web for browsing, finding, reading or downloading information from the internet. However, when considering sending a credit card number over the internet, they are reluctant to do it, even if they are told that the transfer is secured. This is because many media expose bad news about the internet security, although security technology for the internet exists and good enough for protecting transactions via the internet, it requires the customer and merchant to trust each other: an undesirable requirement even in face-to-face transactions, and across the internet it admits unacceptable risks[19].

Secure payment systems are critical to the success of E-commerce. There are four essential security requirements for safe electronic payments (Authentication, Validity, Encryption, Integrity and Non-repudiation).

We will discuss later on here after a brief history of the security protocols, three famous protocols: SSL, SET and 3d-secure and other modified models of 3d-Secure protocol.

A. History of Security Protocols

iKP which was in usage beginning with mid-1996, is actually the ancestor of SET. iKP is known for the longevity,

security and the simplicity of the connection mechanism which made its experience to be unique. SET appeared at the initiative of VISA and MasterCard, in order to satisfy other needs that iKP did, such as: information confidentiality (both the card owner and the seller had to be authenticated in order to protect all parties were involved), independency from other protocols, platforms and operating systems, etc. [2].

However after 2000 new ideas for other protocols much better than SET, started to appear. Moreover since SET proved to be somehow a failure, especially because the actions of the seller were relatively complex. In fact there should have been established more communications with the customer, with the bank and with the payment gateway. Seeing this lack of interest, a new payment scheme was created at the initiatives of Visa.

Compared to SET, 3D-Secure answers a simpler scheme and allows the integration of a much easier usage for the seller and the buyer. Most responsibilities are now transferred to banks. The main innovation in terms of security is the introduction of SSL/ TLS. TLS (Transport Layer Security) is the IETF version of SSL. At the beginning 3D-Secure was called 3D-SSL.

Nowadays it is in general use (starting with 1st of March 2003) and it is supported by Visa, MasterCard, American Express, etc. [2].

B. SSL protocol:

The SSL protocol, widely deployed today on the internet, has helped create a basic level of security sufficient for some hearty souls to begin conducting business over the Web. SSL is implemented in most major Web browsers used by consumers, as well as in merchant server software, which supports the seller's virtual storefront in cyberspace. Hundreds of millions of dollars are already changing hands when cybershoppers enter their credit card numbers on Web pages secured with SSL technology.

SSL is implemented in all popular browsers and web servers. Furthermore, it is the basis of the Transport Layer Security (TLS) protocol. In this sense, SSL provides a secure channel between the consumer and the merchant for exchanging payment information. This means any data sent through this channel is encrypted, so that no one other than these two parties will be able to read it. In other words, SSL can give us confidential communications, it also introduces huge risks:

- The cardholder is protected from eavesdroppers but not from the merchant. Some merchants are dishonest: pornographers have charged more than advertised price, expecting their customers to be too embarrassed to complain. Some others are just hackers who put up a snazzy illegal Web site and profess to be the XYZ Corp., or impersonate the XYZ Corp. and collecting credit card numbers for personal use.
- The merchant has not protected from dishonest customers who supply an invalid credit card number or who claim a refund from their bank without cause. Contrary to

popular belief, it is not the cardholder but the merchant who has the most to lose from fraud. Legislation in most countries protects the consumer [20].

C. SET Protocol

Visa and MasterCard and a consortium of 11 technology companies made a promise to banks, merchants, and consumers: they would make the internet safe for credit card transactions and send electronic commerce revenues skyward. With great fanfare, they introduced the Secure Electronic Transaction protocol for processing online credit card purchases [9]. SET is an open standard for encryption and security specification for credit card transactions on the internet [18]. The SET is a set of security protocols and formats that main section are application protocol and payment protocol. The electronic commerce parties based on SET protocols can be illustrated as Fig. 2. [14]

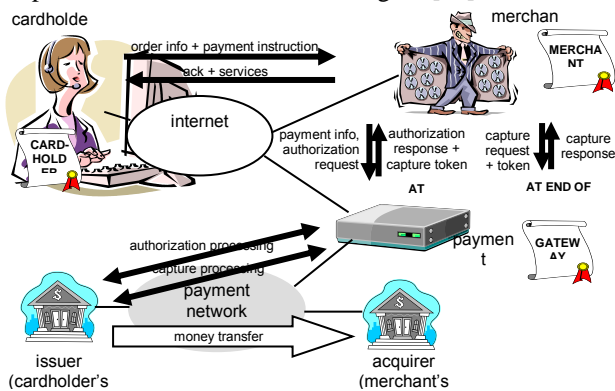


Fig 2: Parties in Set protocol

Key Technologies of SET:

- Confidentiality of information: DES.
- Integrity of data: RSA digital signatures with SHA-1 hash codes.
- Cardholder account authentication: X.509v3 digital certificates with RSA signatures.
- Merchant authentication: X.509v3 digital certificates with RSA signatures.
- Privacy: separation of order and payment information using dual signatures.

Dual Signatures:

An important innovation introduced in SET; the dual Signature. The purpose of the dual signature is the same as the standard electronic signature: to guarantee the authentication and integrity of data. In this case, the customer wants to send the order information (OI) to the merchant and the payment information (PI) to the bank. The merchant does not need to know the customer's credit card's number, and the bank does not need to know the details of the customer's order. The customer is afforded extra protection in terms of privacy by keeping these two items separate. However, the two items must be linked in a way that can be used to resolve disputes if necessary. The link is needed so that the customer can prove that this payment is intended for this order and not for some other goods and

service [15]. So, Dual Signature links two messages securely but allows only one party to read each as shown in Fig. 3 .

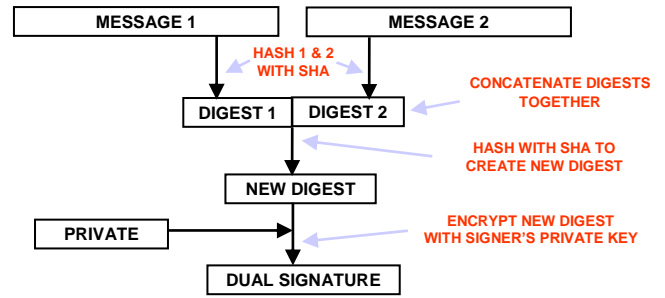


Fig. 3 Dual Signature: Links two messages securely but allows only one party to read each.

Fig.4 shows the model of dual signature. When the dual signature is constructed, it gets the hash of the concatenated hashes of OI (Order Information) and PI (Payment Information) as inputs. The dual signature is the encrypted MD (with the customer's secret key) of the concatenated MD's of PI and OI. The dual signature is sent to both the merchant and the bank. The protocol arranges for the merchant to see the MD of the PI without seeing the PI itself, and the bank sees the MD of the OI but not the OI itself. The dual signature can be verified using the MD of the OI or PI. It doesn't require the OI or PI itself. Its MD does not reveal the content of the OI or PI, and thus privacy is preserved. We can summarize these steps as follows:

1. Take the hash (SHA-1) of the payment and order information.
 2. These two hash values are concatenated [H(PI) || H(OI)] and then the result is hashed.
 3. Customer encrypts the final hash with a private key creating the dual signature.
- $DS = E_{KR_c} [H(H(PI) || H(OI))]$.

$$DS = E_{KR_c} [H(H(PI) || H(OI))]$$

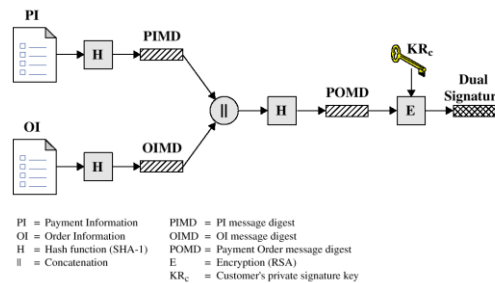


Fig. 4 Dual Signature Model [15]

4. The merchant request payment authorization.
5. The merchant confirm the order.
6. The merchant provides the goods or service.
7. The merchant requests payments.

SET has many merits:

- SET has provided merchant protective method, cost-cutting and enough security for the electronic payment. It makes the business exempted from the online fraud.
- As for the consumer, SET has guaranteed validity of online merchant as credit card number of cardholder will not be stolen. SET keeps more secrets for the consumer to improve the satisfaction of their on-line shopping experience.
- SET helps the bank and the credit card company to expand the service to more broad space – internet. And it lowers the probability of credit card on-line fraud.
- Therefore SET seems more competitive than other online payment method.
- SET has defined interface for all quarters of online transaction so that a system can be built on the products made by the different manufacturers.

Although SET has been widely used in the electronic payment area and has gained more attention from the electronic commerce promoter, the SET transaction mode can not be used on the B2B business model but B2C model only. Even for B2C model, its application is also limited. For example, it can only be applied in some types of card payment service. Its deficiencies mainly display on following aspects [7]:

1. DES algorithm and the RSA algorithm are used in SET protocol to carry on the encryption and the decryption process. SET protocol use DES as symmetrical encryption algorithm. However, DES was no longer a safe algorithm right now. Therefore, DES should be replaced by more intensive and safer algorithm. Moreover, along with the development of processing speed and storage efficiency enhancement of the computer, the algorithm will be cracked successively. It is necessary to improve the extendibility of encryption service.
2. SET protocol is huge and complex in the application process. In a typical SET transaction process, the digital certificates need to be confirmed 9 times, transmitted 7 times; the digital signature need be confirmed 6 times, and 5 times signature, 4 symmetrical encryptions and 4 asymmetrical encryptions are carried out. SET protocol involves many entities such as customers, merchants and banks. All of them need to modify their systems to embed interoperability. As the SET requests installment software in the network of bank, on the business server and PC of the customer and it also need to provide certificates to all quarters, so running cost of the SET is rather high.
3. The protocol cannot prove transactions which are done by the user who signs the certificate. The protocol is unable to protect cardholder and business since the signature received finally in the protocol is not to confirm the content of the transaction but an authentication code. If cardholders and trade companies have the dispute cannot provide alone the evidence to prove its transaction between themselves and the banks.
4. SET protocol specification has not mentioned how to store or destroy this kind of data safely after business processes complete and whether the data should be stored in

the computer of the consumers, or the online store, or in the receipt bank. This kind of Vulnerability possibly will cause these data later under the latent attack.

5. SET protocol has not considered the fairness of transaction individual. Credit card information of cardholder retransmitted through online merchant, although has been undergone the encryption, still could be known by the merchants what the cardholder has bought. This process has not provided anonymous to consumers, consequently it is a serious potential danger.

6. In the document of transaction, time is the especially important information. In the written contract, the date when the document signs and signature are equally crucial content and should be prevented from forge and the distortion. On the other hand, it is easy to changes the timer of some document on the computer. For that reason, the corresponding security measure in the electronic transaction process should be taken to protect the safety of the date and time information of the document.

Meanwhile it could prevent lawsuit from transaction denial thereafter. Although there are some drawbacks in the SET protocol, it is still the most standard and the safest in the present electronic commerce security protocol and the international standard of the security electron payment [20].

D. 3-D Secure Protocol.

The three-Domain Secure (3-D Secure) model of VISA provides the issuers with the ability to authenticate cardholders during an online purchase. This reduces the fraudulent use of credit cards and increases traceability of the transaction. The model divides the payment system into: Issuer Domain, Acquirer Domain and Interoperability Domain [8], as shown in fig. 5.

- The issuer domain is integrated by the Cardholder, a Visa member financial institution (Issuer) and a VISA component Access Control Server (ACS). This domain is responsible for managing the enrolment of their cardholders in the service and for authenticating cardholders during online purchases by means of ACS.
- The Acquirer domain is integrated by Merchant, a VISA financial institution (acquirer) and a VISA component Merchant Server Plug-in (MPI). This domain is responsible for defining the procedures to ensure that merchants participating in the internet transactions are operating under a merchant agreement with the Acquirer, and providing the transaction processing for authenticated transactions by means of MPI.
- The Interoperability Domain is integrated by Visa Directory Server (DS) and Authentication History Server (AHS). The Visa directory Server handles all the communication between Merchant and the appropriate ACS in the process of request if the payment authentication is available. AHS stores the messages from the ACS for each attempted payment authentication and could be used by acquirers and issuers in case of disputes.

The following figure represents the Domain model of VISA and the principal flows in the payment protocol.

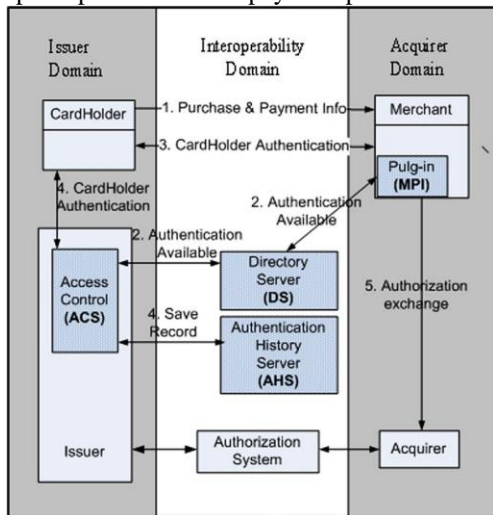


Fig. 5: Domain model of VISA and the principal flows in the payment protocol [8].

The payment protocol in 3-D Secure

Principal Messages

1. VEReq – Message from MPI to the DS or from DS to the ACS, asking whether authentication is available for a particular card number.
2. VERes – Message from the ACS or the DS, telling the MPI whether authentication is available.
3. PAREq – Message request sent from the MPI to the ACS (via the cardholder browser), to issuer to authenticate its cardholder.
4. PAREs – Message formatted, digitally signed, and sent from the ACS to the MPI (via the cardholder browser) providing the results of the issuer’s 3-D Secure cardholder authentication.

Flows of messages

1. First, the cardholder indicates the decision to buy, sending the purchases and payment info at this moment, MPI software is activated.
2. The MPI sends a message (VEReq) to the DS to determine whether authentication services are available for the cardholder.
 - If the cardholder is enrolled and authentication is available, the response message (VERes) instructs the MPI on how to contact the ACS (protocol continues with step 3).
 - If the account number of the cardholder falls outside of participating card ranges, the merchant proceeds with a standard authorization request.
3. The MPI sends an authentication request (PAREq) to the ACS. This is usually sent via the cardholder browser.
4. The ACS authenticates the cardholder by causing an authentication dialog to be displayed to the cardholder asking

for a password, or by some other authentication method, such as a Visa chip card. The ACS formats and digitally signs the authentication response (PAREs), then returns it to the MPI.

5. If the authentication response indicates successful authentication, the merchant forwards an authorization request with the requisite data to its acquirer for submission into an authorization system.[8]

E. Intermediary-3D Secure

Mildrey Carbonell & elt. [8] proposed a multiparty electronic commerce protocol in which the intermediary plays the role of a payment mediator. This intermediary helps the customer to make purchases and payments with many providers simultaneously as a single payment transaction. They proposed model decreases the number of customer operations in the traditional multiparty payment process. This optimization in the payment process for this kind of multiparty scenarios is particularly interesting when we consider the devices which have some resources constraint (computational or connectivity), this is the case of portable devices. Also, in the secure infrastructure proposed, is not assumed to have strong trusting restrictions in the intermediary entity (i.e. not need to be a TTP) which implies a more flexible scenario. In addition, they adapted the 3D Secure_ payment protocol, using their intermediary, to offer the possibility of making secure payment with multiple providers that not need to be enrolled in VISA 3D Secure. In this multiparty electronic commerce model, the intermediary plays the role of payment mediator between one customer and many providers. Here, the customer delegates the multipayment transactions to the intermediary and creates a single secure transaction between customer and his providers. This model is a secure solution in which the customer creates a short-term certificate for the intermediary as authorization credential to forward and distribute the payment info. This will be used in the distribution process to create evidence of the intermediary’s participation. Also by this means, the provider can obtain the customer’s authentication and assurance of purchase integrity. Unlike to other secure solutions in e-commerce models with intermediary, in this secure solution the intermediary is not represented as a trusted entity (is not a TTP) [8].

IV. DISCUSSION

Since 1990s a lot of Security protocols appeared but only a few of them succeeded and became widely implemented. Among the most successful are SSL and SET. Secure Socket Layer protocol (SSL) is used by the vast majority of internet secure transactions. SSL is implemented in all popular browsers and web servers. It was originally designed by Netscape. It was developed to provide encryption and authentication between a web client and a web server. Furthermore, it is the basis of the Transport Layer Security (TLS) protocol.

Secure Electronic Transactions protocol (SET) is another protocol competing with SSL. In E-Commerce whether with SSL or SET, usually uses payment credit and debit card infrastructure. Here we try to answer the big question: which to use SSL OR SET protocol?.

The three major players in this issue is: customers, merchants and financial institutions(usually banks).

We have seen that SSL provides security for communication between the first two players (the customer and the merchant), while SET provides security for communication among all three players. Here we must state some facts about SSL, which are: SSL is the basis of the TLS. Also SSL and TLS are not limited to web applications. In fact, they can be used for authentication and data encryption in IMAP mail access. Furthermore, SSL can be seen as a layer between the application layer and the transport layer. On the sender side, it receives data (for example http messages) from the application layer and encrypts it before directing the encrypted data to a TCP socket. The opposite happens at the receiver side.

SSL is popular today. It enabled servers and browsers provide a popular platform for card transactions. In spite of that, SSL was not developed specifically for card payment, but instead for generic secure communication between a client and a server.

The generic design of SSL may cause problems. For example, by using SSL we can authenticate the customer and the merchant, but we can't be sure whether the merchant is authorized to accept payment, nor whether the customer is authorized to pay money. SSL also doesn't tie a client to a specific card. For these reasons we need a protocol that handles authentication and authorization for card payments transactions. The protocol that could do that was the SET protocol.

SET was developed in 1996 by Visa, MasterCard, Microsoft, Netscape, IBM among others. This protocol was designed specifically to secure card payment transactions over the internet. It encrypts payment related messages. SET can't be used for general purposes like encrypting arbitrary text of images. SET involves all three players in E-payment. In SET all three players must have certificates.

The customer's and merchant's certificates are issued by their banks in order to assure that they are permitted to make/receive payments by cards. In a SET transaction, the customers card number is passed to the merchant's bank. This number is never seen by the merchant as plaintext.

SET beats SSL in secure issues since it has the following properties:

- All players must hold trusted certificates.
- All parties are authenticated.
- SET provides privacy, merchant will never see the customer's card number.
- SET provides data integrity.
- SET provides customer non-repudiation guarantee.
- SET provides customer and merchant authorization.

But in the other side, SET is not easy to implement and SET requires the customer to install an e-wallet. It is expensive to integrate with legacy applications.

SET is safe since it addresses all the parties involved in typical credit card transactions: consumers, merchants, and the banks. Besides the interoperability problem, it has difficulties to spread since it needs all the participants to have

some part of the software, even very expensive hardware. It may be clearly in the interests of the credit card companies and banks, but it looks quite different from the perspective of merchants and consumers. In order to process SET transactions, the merchants have to spend several million dollars in equipment and services when they already have what are arguably sufficient security provisions in SSL. To consumers, they have to install software also.

SET can work in Real Time or be a store and forward transfer, and is industry backed by the major credit card companies and banks. Its transaction can be accomplished over the WEB or via email. It provides confidentiality, integrity, authentication, and, or non-repudiation.

SET is a very comprehensive and very complicated security protocol. It has to be simplified to be adopted by every parties involved in E-commerce transaction.

In my opinion, for merchants to build trust, they should adapt the SET protocol, in spite of its heavy cost.

V. CONCLUSION

Electronic commerce is now one of the widest applications in internet since it helps businesses to expand their marketing strategy and to reduce their costs. This growth has motivated the development of research to improve electronic services. Security, as one of these research topics, constitutes a critical point in the implementation of new business models because the process of traditional business such as paper-based contracts, personal purchases, etc. must be adapted to flows of information inside an unreliable network like the internet. So, payment should be the process with the highest security level in e-commerce operations because it is the step where the customer legally ends the business by making the money transference.

Many secure electronic payment solutions have been proposed. Some of them describe online payment with a cash payment model, like e-Cash, DigiCash, NetCash, and Cybercash. Others, such as NetBill, NetCheque and BankNet, present a cheque payment model. And, in a card payment schema, open solutions such as iKP and SET have been developed as a standard of secure payment. iKP and SET were not widely used in the internet but they constitute a starting point in the development of secure payment solutions. Today, the most popular solution in the card payment schema is the 3-D Secure protocol (3-D Secure) developed by VISA and MasterCard, which is based on the ideas of iKP and SET. This protocol provides the card issuer with the ability of authenticating its cardholders during an

online purchase. Given that VISA has licensed this protocol and that many vendor communities use it, 3D Secure is considered a standard for authenticated payment [8]. This paper focused on SET and, 3D Secure protocols.

Electronic payment solutions are mostly focused on traditional two party business models. However, many business models involve some intermediary entities to help negotiation. Mildrey Carbonell & elt. [8] proposed a multiparty electronic commerce protocol in which the intermediary plays the role of a payment mediator. This intermediary helps the customer to make purchases and payments with many providers simultaneously as a single payment transaction. Even though, these proposed models solved many problems of previous one, they still suffer many gaps in the payment process.

We expect to see other advanced models in the near future that overcome all the disadvantages of the previous protocols and models.

REFERENCES

- [1] Adam Jolly, "The Secure Online Business", Great Britain and the United States- Kogan Page Limited 2003, pp: 93-118.
- [2] Be l l a r e , M . , G a r a y , J . , H a u s e r , R . , Journal of Selected Areas in Communications, Vol. 18, No.4, 2000.
- [3] Donal O.Mahony, Michael Peirce Hitesh Tewari, "Electronic Payment Systems for E-Commerce", Artech House computer security series-Boston 2001, Second Edition, pp: 19-69
- [4] E. Harrison McKnight and Norman L. Chervany., "What Trust Means in E-Commerce customer Relationships: An Interdisciplinary Conceptual Typology", International Journal of Electronic Commerce , 2001-2002, p. 35-59.
- [5] "eCommerce Trust Study", research report, Cheskin and Studio Archetype/Sapient. Materials of Dagstuhl Seminar,1999.
- [6] Hawker, A., "Security and Controls in Information Systems", London, Routledge, 2000.
- [7] M Franklin, M Yang, "Towards Provably Secure Efficient Electronic Cash," ReportCUCS-018-92. Columbia University Department of Computer Science, 2005.
- [8] Mildrey Carbonella, Jose' Mari'a Sierraa, Javier Lopezb, "Secure Multiparty Payment with an Intermediary Entity", computers & security 28 (2009) 289 - 30.
- [9] Nikki Goth Itoi., "PROMISES, PROMISES What ever happened to SET", available at: <http://www.herring.com/mag/issue51/promises.html>.
- [10] NIST (1995), "The Data Encryption Standard": An Update, <http://csrc.nist.gov/publications/nistbul/csl95-02.txt>.
- [11] PETER C. CHAPIN, CH. SKALKKA, and X. SEAN WANG, "Authorization in Trust Management: Features and Foundations", ACM Computing Surveys, Vol. 40, No. 3, Article 9, August 2008, pp: 9.1-9.48.
- [12] Pradnya B. Rane , Dr. B.B.Meshram, "Transaction Security for E-commerce Application", International Journal of Electronics and Computer Science Engineering. Available Online at: www.ijecse.org ISSN- 2277-1956. pp: 1720-1726.
- [13] Rivest, R., A. Shamir and L. Adleman, "A Method for Obtaining Digital Signatures and a Public Key Cryptosystem", Communication of the ACM, vol 21, pp.120-128, 1978.
- [14] S Lu, S Smolka, "Model checking the secure electronic transaction (SET) protocol," Proceedings of the 7th International Symposium on Modeling Analysis and simulation of Computer and Telecommunication Systems, 1999:358-364.
- [15] S. William, "Cryptography and Network Security: Principles and Practice", 2nd edition, Prentice-Hall, Inc., 1999 pp553- 554.
- [16] Stuart Feldman, "The Changing Face of E-Commerce: Extending the Boundaries of the Possible", IEEE INTERNET COMPUTING, MAY • JUNE 2000, p.:82-83.
- [17] Vijay Ahuja, "Building Trust in Electronic Commerce", IEEE/2000, pp:61-63.
- [18] Visa and MasterCard, "SET Secure Electronic Transaction", Book2, Programmer's Guide,1997.5.
- [19] Ya n g L i , Y u n W a n g , "Secure Electronic Transaction", <http://islab.oregonstate.edu>.
- [20] Z. Boping, Sh. Shiyu, "An Improved SET Protocol", Proceedings of the 2009 International Symposium on Information Processing (ISIP'09), Huangshan, P. R. China, August 21-23, 2009, pp. 267-272, ISBN 978-952-5726-02-2 (Print), 978-952-5726-03-9 (CD-ROM).

Towards Building Novel Educational System for School Students Using Smart Phones and QR Codes

Yousef A. Eyadat

Department of Curriculum and Instruction
Yarmouk University
Irbid, Jordan
eyadat@yu.edu.jo

Reem A. Wahsheh

Department of Curriculum and Instruction
Yarmouk University
Irbid, Jordan
reem_ahmad_wahsha@yahoo.com

Yarub A. Wahsheh

Department of Network Engineering and Security
Jordan University of Science and Technology
Irbid, Jordan
y_wahsha@yahoo.com

Abstract—With the increase of using smart phone devices, these devices can play important roles in many life fields, education is one example of these fields. Smart phones can be used to read QR codes, which can link physical objects into electronic resources. In this project we propose a novel educational system for school students that use QR codes and smart phone devices or tablet computers. We aim to link the physical school books to additional electronic resources allowing students to reach educational games, multimedia resources and online experiments. Our educational system allows communication between students and teachers and even between parents and teachers through educational server. Also the proposed system gives the students' parents the ability to access all their children information using one account.

Keywords— QR Codes, Smart phones, E-Learning; Educational System

I. INTRODUCTION

Quick Response (QR) code is a two dimensional barcode that stores data in two dimensions and can be read using an imaging device, such as: smart phone, tablet computer or specific scanning device [1]. Nowadays smart phones are becoming more and more popular and important, they offer a lot of features that make our life much easier than ever before. These features include voice communication, accessing the Internet any time and everywhere, using digital cameras and viewing multimedia resources. One more important feature is that smart phones can be used as scanning devices for QR codes.

QR codes allow the reader application to complete an action. It links physical objects (books, posters or advertisements) to specific electronic encoded data (web page

address, email, SMS message or text information). The reader application can decode this data and read some text, get specific parameters, redirect browsers to a specific web page or use contact information [1]. QR codes can contain any type of information with a specific size. According to QR code standard [2] the encoded data can be numeric, alphanumeric, binary and Kangi data, with a size up to 4296 characters for alphanumeric data. Figure (1) shows a QR code that contains a link to Yarmouk University website.



Fig. 1. QR code contains a link to Yarmouk University website.

QR codes are increasingly used to cover various fields, such as: product tracking, item identification, contact information and general marketing [1]. One interesting field is using QR codes with mobile phones in the class room environment for educational purposes. QR codes can be used to provide students with just in time resources [3] [4].

The rest of this paper is organized as follows: Section two presents a brief of the related studies. Section three explores the proposed educational system importance. System structure and algorithm are shown in section four. Section five explores the advantages and challenges of the proposed system and finally section six presents the conclusion and future work.

II. RELATED WORK

The study of [3] aims to improve college classes to use smart phones and QR codes. Improvements include developing a system that allows college students to answer questions about the class and send their comments or suggestions to their teachers and classmates during classes. The system was developed using smart phones and QR codes in the middle of each class.

The study of [4] explores the evolution of smart phones into a powerful tool for education, providing a literature review on the usage of smart phones in higher education for both professors and students and how smart phones can be useful inside and outside the class room.

The study of [5] is a report study on the potential of using QR codes in learning and education, providing several scenarios of QR code usage in presentations and class room feedback, with survey results for university students asking them about QR codes and smart phones usage in education.

Objects identification system to help blind and visually impaired people was developed in [6]. The study propose objects identification system for blind people using QR codes and smart phones, the study shows how QR code can be used usefully in real-time interaction with different environments. Where in [7] a novel educational information system for Holy Quran was developed using QR codes, linking hard copies of Holy Quran with electronic resources for audio, translation and interpretation for Holy Quran learning process.

III. PROPOSED SYSTEM IMPORTANCE

In this part we propose the importance of building novel educational system for school students depending on QR codes and smart phones (or tablet computers) that are connected to the Internet. Our goal is to improve the educational process in general and allow school students to access online materials including multimedia resources and educational games. Also the proposed system allows the students and their parents to contact with school teachers through educational server.

To determine the importance of building such a system we did a survey for school students. The survey contained nine questions in Arabic language, with a sample size of 180 students who are between 12 and 17 years old. The survey was printed and applied in Jordan National Schools (Irbid – Jordan). In our survey we ask students about their usage of smart phones and QR codes, Table 1 shows the survey results.

TABLE I. SURVEY RESULTS

Question	Students who answered yes - percentage	Students who answered no - percentage
Do you have a smart phone device?	122 – 67.7%	58 – 32.2%
Do you have Internet connection on your mobile device anytime and everywhere? (of those who have smart phones)	63 - 51.6%	59 – 48.3%
Have you heard about Smart phones usage in educational field?	95 – 52.7%	85 – 47.2%
Have you heard about QR code?	70 – 38.8%	110 – 61.1%
Do you use the URL links printed in school books to get more information about lessons?	54 - 30%	126 - 70%
Do you use the internet to access extra resources and experiments?	54 - 30%	126 - 70%
Do you prefer using multimedia resources in the educational process?	117 - 65%	63 - 35%
Do you prefer to have communication with your teachers outside class rooms?	117 - 65%	63 - 35%
Do you check your marks online on the e-learning system?	98 – 54.4%	82 – 45.5%

Survey results showed that the majority of the students use smart phones and Internet but they are not aware of what QR code is. It is important to use this wide spread of smart phone devices that are connected to the Internet in a beneficial way. The students can gain great benefits with electronic resources, educational games, multimedia resources and online experiments. Survey results showed that the usage of smart phones can play important roles developing the educational process.

IV. PROPOSED SYSTEM STRUCTURE AND ALGORITHM

Our proposed system contains three main components: QR code label, Smart phone and Educational server.

QR codes are used to link physical papers of student books with the corresponding electronic materials and educational

resources. Smart phones are used as scanning devices; they read the QR code contents and use them to connect to specific educational server that contains the educational resources.

In order to manage these operations we propose a specific structure to encode data in QR codes, this structure is important to define common well known rules so the scanner devices (mainly smart phones) can handle these data and use them efficiently. Figure (2) shows our proposed system components.

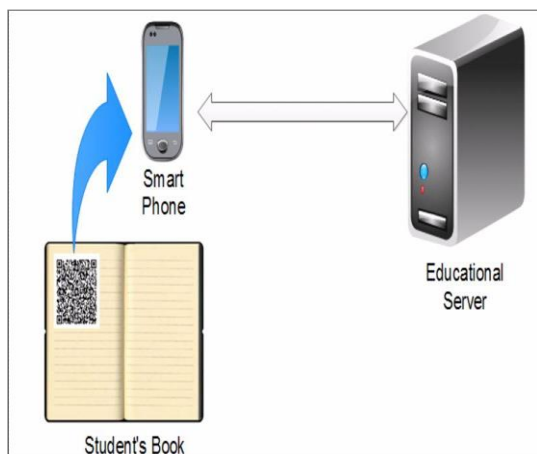


Fig. 2. Main components of the proposed system.

Our proposed structure uses the concept of Extensible Markup Language (XML) tags. XML is a markup language that aims to define data encoding rules in a format that is both human-readable and machine-readable [8].

For student books we need to add QR code label to each lesson page. QR codes will contain specific parameters that give full definition of the lesson such as: grade, course subject, unit number, lesson number, title and even extra URLs that support the materials. Also QR codes should contain educational server information (domain name or IP address). This information will be encoded in XML; two main tags are used to encode these parameters; lesson info tag and server info tag. Figure (3) shows an example of data encoding in tags.

```
<lesson_info>
  <grade>ninth</grade>
  <course>chemistry</course>
  <unit>5</unit>
  <lesson>2</lesson>
  <page>56</page>
</lesson_info>
<server_info>
  <domain_name>www.example.com</domain_name>
</server_info>
```

A square QR code is positioned to the right of the XML code block, representing the encoded data.

Fig. 3. Example of data encoding in tags.

Our proposed system algorithm steps are:

- Encode the parameters (as defined in system structure) of each lesson in QR code and print this QR code on the corresponding student book paper.
- Students and parents will use specific reader application that is installed on their smart phone devices to scan QR codes.
- Reader application is responsible for data parsing and understanding the encoded parameters.
- Reader application will use server information to establish a connection with the corresponding educational server and will send the parameters that identify the lesson.
- Educational server will handle student accounts with user name and password and these accounts will contain all the needed resources, educational games, multimedia resources, marks, experiments, extra materials and teachers contact information such as email or phone number.
- Educational servers may handle parents' accounts that contain full information about their children and contact information with teachers.
- Educational server will regulate the communication process between students and teachers and between parents and teachers. This can be done using messaging system running on the educational server, or simply the server will provide information about the subject teacher including his/her name and contact information such as phone number or email.

Through the connection between the smart phone and the educational server, students will be able to use the educational server services. Figure (4) shows the proposed educational system.

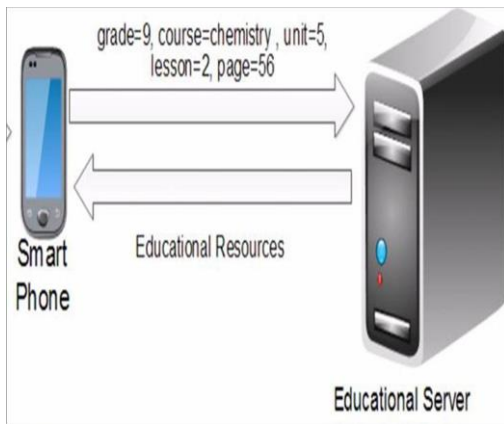


Fig. 4. The proposed educational system.

V. SYSTEM ADVANTAGES AND IMPLEMENTATION

A. Advantages and challenges

Our proposed system offers important features and services to the educational process. Main advantages are:

- It is easy to implement.
- There is no need to type URLs or key words to access extra information.
- All educational resources are collected in one place (educational server).
- It solves the problems of danger or costly experiments by supporting multimedia resources.
- QR codes are free to use and easy to generate.
- It offers (student - teacher) and even (parents – teacher) communication.

Although the proposed system offers important services, barriers and challenges appear such as:

- Not all students have smart phones. According to our survey 32.2% of the overall number of students does not have smart phone devices, especially for those who are less than 15 years old.
- Not all students have Internet connection on their smart phones and not all schools are covered with Wi-Fi Internet connections. Our proposed system needs Internet connection between the educational server and the student who use smart phone. According to our

survey 48.3% of the overall number of students who have smart phones do not have Internet connections.

B. Implementation and Experiments

The algorithm was implemented using Java server and Android mobile application, figure 5 shows the main interface of the application.

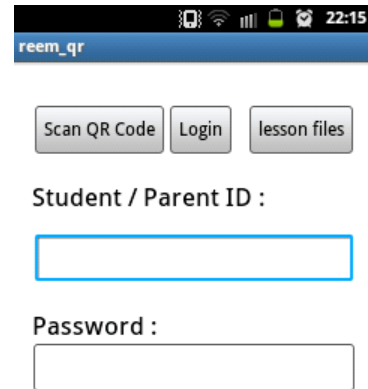


Fig. 5. The main interface of the android mobile application.

The application has 3 main buttons and two text fields. The first button allows the user to scan a QR code; figure 6 shows the contents of QR code that was used as experiment.

```
<lesson_info>
<grade>11</grade>
<course>computer</course>
<page>13</page>
</lesson_info>
<server_info>
<ip>192.168.0.101</ip>
</server_info>
```

Fig. 6. The QR code contents that was used in the experiment.

We have used an online QR code generator to encode these contents, QR stuff website [9] which offers free QR code generation service. The generated QR code is shown in figure7.



Fig. 7. The QR code that was used in the experiment.

Using the first button “Scan QR code” the user can read QR code contents using his smart phone camera. This step is shown in figure 8 and figure 9.

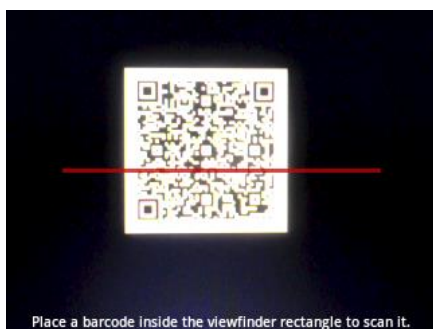


Fig. 8. Scanning the QR code.

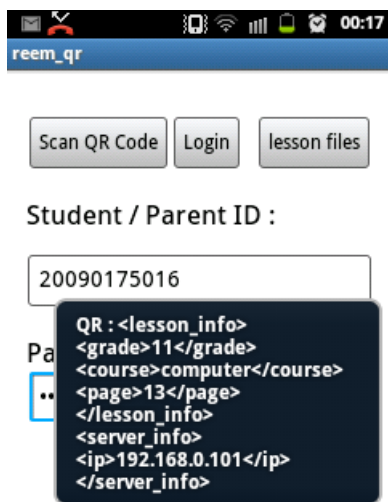


Fig. 9. Reading QR code contents.

The users are asked to enter their student ID (or parent ID) and login to the educational server, parents can access information about all their children in one account.

The educational server will verify user id and password then provide the user with the needed information such as marks, exams, external links that contain more educational resources, multimedia experiments and educational games. Also it will provide teachers' contact information or special messaging system to handle student-teacher and parent-teacher. Figure 9 shows an example of extra resources provided by the educational server.



Fig. 10. Reading QR code contents.

The educational server can be used to access the E-learning system that is used by the Jordanian ministry of education [10]. The E-learning system is used to access students' accounts, marks and courses.

VI. CONCLUSION

In this paper we proposed novel educational system for school students using QR codes and smart phones. We have applied a survey for school students to determine the importance of developing such system, survey results showed that the majority of school students have smart phones but they do not use them in education. Our proposed system offers beneficial use of smart phones in the educational process by defining specific structure for data encoded in QR codes, then using these QR codes to link physical books to electronic educational server which provide students with extra resources, multimedia experiments, educational games, and many other services. Also the system offers great features for students' parents include viewing students' marks and communicate with school teachers.

In the future we are looking to apply this system for schools, take feedback from students, parents and teachers to add more educational services.

REFERENCES

- [1] Wikipedia, Retrieved January, 5, 2015, from http://en.wikipedia.org/wiki/QR_code.
- [2] ISO/IEC 18004, Information technology – Automatic identification and data capture techniques – Bar code symbology – QR Code. 2000.
- [3] Susono, Hitoshi, and Shimomura Tsutomu. Using Mobile Phones and QR Codes for Formative Class Assessment. 2006.
- [4] Yu, Fuxin(Andrew).Mobile/Smart Phone Use in Higher Education, 2001.
- [5] Ramsden, AndY. The use of QR codes in Education: A getting started guide for academics. 2008.
- [6] H. S. Al-Khalifa, “Utilizing QR Code and Mobile Phones for Blinds and Visually Impaired People,” K. Miesenberger et al. (Eds.): ICCHP 2008, LNCS 5105, pp. 1065–1069, 2008.
- [7] H. A. Wahsheh, Y. A. Wahsheh, R. A. Wahsheh, “; Novel educational System for holy Quran using QR codes, ” Proceedings of Al-Zaytona University International Engineering Conference on Sustainability in Design an Innovation ' 2014 May 13-15; Amman – Jordan.
- [8] Wikipedia, Retrieved January, 5, 2015, from <http://en.wikipedia.org/wiki/XML>.
- [9] QR Stuff, Retrieved January, 5, 2015, from <http://en.wikipedia.org/wiki/XML>.
- [10] EduWave, Retrieved January, 5, 2015, from <http://eduwave.elearning.jo/Eduwave/ElearningMe.aspx> .

Cross-Language Name Matching for Data Fusion in Linked Open Data

Ziad F. Torkey

Computer Science

Arab Academy for Science and Technology

Cairo, Egypt

ziadtorky@gmail.com

Emad Elabd

Faculty of Computers and Information

Menoufia University

Menoufia, Egypt

emadqap@gmail.com

Mostafa Abdelazem

Faculty of Computer Science

Arab Academy for Science and Technology

Cairo, Egypt

melbaqary@gmail.com

Abstract -- *Data quality and accuracy affects the success of data integration in Linked Open Data (LOD). The main goal of data fusion is to represent each real-world entity once on the Web. Data inaccuracy problems exist due to misspelling and a wide range of typographical differences mainly in non-Latin languages, those problems become more complicated when a person is identified by a name, and this name can be presented differently in same/different languages. Up to author's knowledge, the previous approaches which supported Arabic person names are not designed to work with LOD. This paper proposes a framework that uses person names as matching criteria from cross-language LOD Datasets. The proposed framework has substantial improvements in matching results compared to state of the art framework of matching techniques with better matching rate which exceed 6% in precision and 6% in recall.*

Key words—*Data fusion; ontology Alignment; duplicate detection; linked open data; semantic web.*

I. INTRODUCTION

Information technology plays an important role in today's IT based economy. Many industries and systems depend on the accuracy of data to carry out operations [1]. In the typical Web (Web 2.0), there are links between documents and the relationship between any linked documents is implicit. Sometimes the information in the web is redundant and the same data has multiple representations [2].

Due to the redundancy of real-world entities over the web, the idea of URI (Unified Resource Identity) was presented in the Linked Open Data (LOD) [2]. Linked Open Data (Web 3.0) represents the same real-world object into unique identity and consistent representation [3].

LOD is a collection of Ontology [2] published over the Web to present things uniquely. Ontology are released in the form of resource description framework (RDF). Ontology over the LOD contains millions of RDF triples (subject, predicate and object). LOD elevated links between different datasets/data sources which characterized the relations between things to facilitate browsing for users [1].

Things in different LOD ontology are presented like (Companies, food, persons) as datasets. Dataset producers like Dbpedia [4] publish datasets for different categories of things based on the available data they have. This leads to the problem of presenting the same real-world entity more than once with different data available on each source of data [3].

The quality of the data stored in LOD can have significant cost effect on a system that uses the information to conduct business. Data fusion is needed in LOD applications to enhance the quality of data [3]. The main goal of data fusion is to integrate different data which represent the same real-world objects and the resolution of data conflicts. The quality of data can be affected by many factors including spelling mistakes, errors in data entry and different conventions in storing information. For example, Arabic data has more problems than Latin based language (English, French and German) because of these different conventions [5].

Arabic is the main language for millions of people in twenty Middle East and North African countries [6] [7]. Arabic language has characteristics like absence of capital letters, complex morphology and short vowels [8]. Since Arabic is one of the languages used in the published Datasets, Data fusion is used for matching things presented in different data sources using different languages.

One category of the published datasets in LOD is datasets which present persons information. Datasets are produced by different vendors all over the Web. Person's datasets present all available data that can be found about the person like (Name, Age, Work in, Birthdate, etc.). Names can be used as matching criteria for those persons across different data sources. Since the same person can be presented in different data sources in different languages (English- Arabic), this means they can be matched using his/her name and any other available data for this person.

Names written in English cannot directly be matched to names written in Arabic due to different language script and morphology. We propose in this paper a matching framework that is based on phonetic techniques to match person names across different sources in different languages (English – Arabic) so that a single person is presented once on the LOD.

The rest of the paper is organized as following: Section II demonstrates the problem in name matching across different languages (English – Arabic). Section III presents an overview about the work done in Ontology alignment field. Section IV describes the proposed framework and how it can help in improving the matching results between English and Arabic names. Section V shows the impressive result using the proposed framework and also in comparison to latest frameworks available. Section VI concludes the paper and the future work.

II. PROBLEM DEFINITION

Data fusion is one of the biggest problems in providing a trusted source of data [2]. Data fusion is needed to achieve the main objective of Linked Open Data, which is presenting a real-world entity once with a single unified resource identity (URI).

One of the problems of data redundancy over the LOD is Person's data. Person data can be redundant due to misspelling or different presentation in different languages (English –Arabic) over the LOD [9]. Same person name cannot be matched in two different datasets written in different languages like (English dataset- Arabic dataset). In addition to that datasets can have misspelled person names in both languages which increase the difficulty of fusing person data.

Therefore, the contributions of this paper are significant for many reasons. Firstly it proposes an automated technique that enhances string matching between multiple data sources containing redundant data. Secondly the framework has the ability to matching person data in cross-language (English-Arabic) using person names as matching criteria. Finally the framework preforms person names matching on different ontology in LOD.

III. RELATED WORK

Data fusion is a problem which still needs a lot of work to be done [3] [10] [11]. Work has been accomplished to propose fundamental techniques for string matching [12] [13]. String matching is a classical problem which has been there before known in databases integration [1]. For more than five decades, the traditional database community has discussed this problem and a lot of work has been done [14].

String matching techniques can be grouped into three classes [13]: global versus local, set versus whole string and perfect-sequence versus imperfect-sequence [13], the first class refer to the amount of information the technique needs to classify a pair of strings as a match or no-match, global techniques start with computing information over the string labels in ontology triples before it matches any strings, in local techniques the string pairs are being considered as the only input required, examples of this class are:

- TF-IDF: this technique is based on that two strings are similar if they share a word that is rare in the ontology [15].

- Soft TF-IDF: this technique is based on Jaro Winkler technique which works on words equality rather than exact match.

The second class consists of two sub-classes, perfect-sequence which requires characters in the pair of strings to occur in the same order so it can be considered as match, imperfect-sequence is using the same technique as the perfect-sequence with a relaxing condition based on a threshold, this condition increases the false match's rate, and examples of this class are:

- Jaccard: The number of words in pair of strings which are having common characters divided by the total number of the unique words in each string.
- RWSA (Redundant, Word-by-word, Symmetrical, and Approximate): strings characters are replaced by their Soundex code, there is a match if each word in the shorter string has a weighted edit distance less than a threshold from a word in the longer string [13].

The third class works by finding the overlap between pair of strings, it works better on long strings, examples of this class are:

- Levenstein [12]: the number of substitutions needed to transform one string to another.
- N-gram [16]: string is converted to a set of n-grams, the results are compared using similarity metric.

Many techniques are explored including machine learning. Some frameworks use training data to semi-automatically find an entity matching strategy to solve a match problem. The quality of the computer string matching process is found to be higher than the manually linked record (done by humans) [17]. TAILOR [18] is a flexible record matching toolbox which allows the users to apply different duplicate detection methods on the datasets. BigMatch [19] is a duplicate detection program which is used by the US Census Bureau. If the sizes of the datasets are large, online record linking can be used [10]. FEBRL [20] is one of the tools that perform record linkage/duplicate detection process. FEBRL includes a new approach for improved data cleaning and standardization that support parallelization [21]. FEBRL needs to be installed on a local machine and configure the operating system and

prerequisite software to match FEBRL platform requirements, which is not suitable for use on the web.

DRDAA (Duplicate Record Detection with Arabic adjustment) is the latest Web-based matching framework that supports cross-language string matching [22]. DRDAA has predefined rules which have been set by subject expert matter in the field of Arabic especially Arabic names. DRDAA has the ability to find matching person names from two different data sources from different languages (Arabic-English). DRDAA is based on rules which have been based on human experiment, which means that the framework is limited to expert's knowledge.

Up to our knowledge and experiments with the current available frameworks and tools, most of them does not support the matching of Arabic names, and none of them support matching cross-language names in Linked Open Data.

IV. CROSS-LANGUAGE NAME MATCHING FRAMEWORK

The proposed framework is a Web-based string matching based on person names in cross-language. Cross-language Name Matching Framework is designed and implemented to overcome the missing feature of names matching in Linked Open Data. The architecture of the proposed framework consists of Datasets selection, Names Triples Listing, Data cleaning and standardizing, creating phonetic coding and finally Name matching and linking as shown in Figure 1.

Datasets from cross-languages (Arabic-English) in Linked Open Data are selected as an input for the framework. Triples that contain person's names are used for matching. Data cleansing and standardization is required for insuring the quality of data for matching. The proposed framework provides name matching between Datasets that have redundant data about persons. Datasets can be fused by matching the names of persons using Soundex and ASoundex techniques and creating new triples for Soundex code values. Soundex code triples gets compared for matching and when a match is found a new SameAs triple is created between the two datasets showing the equality between the two entities (persons) in both datasets.

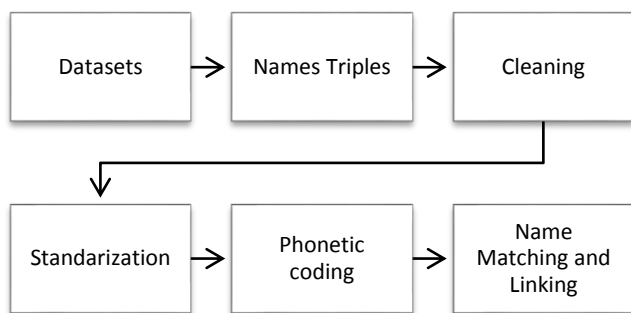


Fig.1. Cross-language Name Matching Architecture

The following sub-sections discuss the framework in details.

A. Data sources

Person’s datasets were selected from FOAF:Person and yago [23] datasets. Two datasets were used as an input for the framework, first in English and second in Arabic. Person datasets contain data like (Name, Age, Birthdate, etc.). Datasets are converted into RDF Graphs [2] so person data is represented in a form of triples as shown in figure 2. Person name triples are selected for cleaning and standardizing.

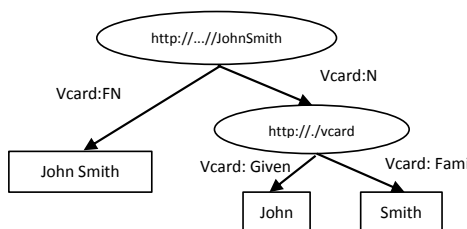


Fig.2. Example on RDF Triples

B. Data Cleaning and Standardizing

The framework uses triples containing person names as an input for cleaning process. Data cleaning is the process of removing all inconsistent or rubbish data like (null triple, “aaaaa”, “_”, “|||”), example on inconsistent data is finding only numbers in name property or finding characters in an integer property like date or age.

In the datasets names might be stored with prefixes. An example for this problem is “prof. ahmed”, these names can be represented in different ontology as “ahmed”, “استاذ

احمد” and all of them represented the same real-world object. After studying number of data samples, sometimes names are written with a prefix like (“Dr. Mostafa”, ”Eng. Magdy”, ”junior”, ”عماد”, ”أ. عماد”, ”السيد طارق”, ”أ. عماد”). Those prefixes were collected in following table:

Table 1: Names prefixes in English and Arabic

English Prefix	Arabic Prefix
Dr.	دكتور
Prof Dr	استاذ دكتور
Prof.	استاذ
Prof	أستاذ
Eng.	.م
Eng	مهندس

Prefixes are removed from names properties in the selected triples so the remaining string in the object is the name without any distraction. Finally the name string is trimmed to remove any unrequired spaces in the string.

One of the problems in person names in Latin based languages is the multiple representation of a word [12]. In Arabic language, the problem may occur in one character like “أ” which can be represented as “أ، آ، إ” and this character using will be based on the pronunciation. Data standardization is used to unify the Arabic Dataset so that misspelling or different pronunciation can be controlled and unified. Set of standardization rules shown in Table 2 are applied on Arabic dataset.

Table 2: Standardization rules

Set of characters	Equivalent character value
ا، آ، إ، ا	ا
ي، ي، ي	ي
ه، ه، ه	ه
و، و، و	و

C. Phonetic coding

Previous approaches worked on aligning ontology in different languages base on translation [24]. Person names cannot be translated, if an Arabic name like “سعيد” (Seed) was translated into English it would be “happy” which is not the same meaning. The proposed framework converts names into phonetic code. This phonetic code should be

equivalent in any language when phonetic technique is used.

Soundex was invented by Russell [25], it is the most common phonetic coding scheme for Latin-based languages, and it is based on replacing characters with phonetic code. ASoundex [26] was introduced later for Arabic language which followed the same pattern of Soundex with a little bit of tweaks.

The proposed framework uses Soundex and ASoundex for creating the phonetic code. String name value found in name, full name or given name property. The selected list of triples is converted into Soundex/ASoundex code and stored as new triple attached to the person URI, later on new Soundex or ASoundex property triples get attached back to the original Graph. This Soundex/ASoundex code property is used as matching criteria between datasets from two different languages (English-Arabic) which may contain redundant data about persons.

Based on the following table the initiation of the Soundex code is created:

Table3: Initiation of Soundex and ASoundex Code

Code	Characters	English phonetic equivalent	Category
1	ب،ف	b,f	Labial
2	خ،ج،ز،س،ص،ظ،ق،ك	k,q,z,s,c,z,j,kh	Guttural and sibilants
3	ت،ث،د،ذ،ض،ط	t,d	Dental
4	ل	l	Long liquid
5	م،ن	m,n	Nasal
6	ر	r	Short liquid

Problems were found in English names like “Charly”, it can be pronounced in Arabic as “شارلي” so that we can consider “Ch” as “ش” in Arabic, but when we ran to a name like “Christen” we found that “Ch” is considered as “ك” in Arabic. Another problem was in an Arabic name like “اسامة”, this name can be written in English as “Osama” or “Usama”, both pronunciations are correct. With using Soundex technique first character gets reserved which mean some times it will be “U” and other times it will be “O”. Some conditions needed to be added to the framework to solve those problems.

D. Name Matching

Soundex code triples from English and Arabic Graphs get compared looking for similarities. When equal code triples are found a new owl:SameAs triple gets created

between the two entities from both Graphs (English – Arabic) as shown in figure 3.

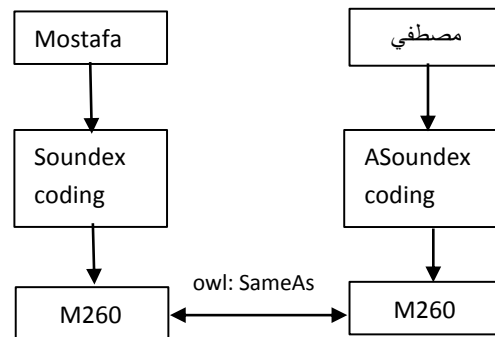


Fig.3. Example of Name Matching using Soundex and ASoundex coding

V. EXPERIMENT AND RESULTS

For testing the framework, English and Arabic datasets were selected to perform number of experiments to check the performance of the framework. Based on the authors’ knowledge, DRDAA framework was chosen for comparing the performance on the proposed framework.

A. Experiment 1: Sample of 100 names triples.

In this experiment we used English and Arabic datasets containing 100 person data. The table below shows the results of precision and recall for this experiment.

Table 4: Experiment 1 results.

Quality metric	Proposed Framework
No. of Entities	100
True Positives (TP)	79
True Negatives (TN)	0
False Positives (FP)	2
False Negative (FN)	19
Precision (TP/(TP+FP))	98%
Recall (TP/(TP+FN))	81%

Experiment 1 results 98% in precision and 81% in recall. After investigating these results, we found that some English names can be pronounced differently if it starts with a special sequence of characters. Examples for those characters are (“C”+”H”+”R”) and (“C”+”H”+”A”). Some tweaks needed to be added to the standard Soundex technique to overcome this problem.

B. Experiment 2: Sample of 100 names triples

In experiment 2, same sample of person data used in experiment 1 were used in experiment 2. The table below shows the results of precision and recall for this experiment after modifying the Soundex technique.

Table 5: Experiment 2 results

Quality metric	Proposed Framework
No. of Entities	100
True Positives (TP)	96
True Negatives (TN)	0
False Positives (FP)	1
False Negative (FN)	3
Precision (TP/(TP+FP))	98 %
Recall (TP/(TP+FN))	96 %

Experiment 2 gave significant results with 98% in precision and 96% in recall.

C. Experiment 3: Comparison between DRDAA and the proposed cross-language name matching framework.

In this comparison, 3 thousand triples of English and Arabic person names extracted from FOAF: Person [27] and yugo [23] datasets were used for testing the framework. Comparing the proposed framework with the latest similar framework for string matching which is DRDAA [22] we found an improvement in the results as shown in Table 6.

Table 6: Comparison between results between DRDAA and proposed framework

Quality metric	DRDAA	Proposed Framework
No. of Entities	3000	3000
True Positives (TP)	2415	2742
True Negatives (TN)	0	0
False Positives (FP)	210	39
False Negative (FN)	375	219
Precision (TP/(TP+FP))	92%	98%
Recall (TP/(TP+FN))	86%	92%

Comparing the proposed framework results with DRDAA framework which is the state of the art in data fusion and record linkage we found that we have the advantage of higher matching rate which exceed 6% in precision and 6% in recall. The proposed framework is fully automated which is an advantage over the DRDAA that is based on

subject expert matter experience and that can be a limitation for this framework. Finally the proposed framework is the only person name matching approach that is available for data fusion in Linked Open Data.

VI. CONCLUSION AND FUTURE WORK

Data fusion is an important step in Ontology alignment. In this paper, a web-based framework for cross-language name matching in LOD is proposed with enhanced phonetic technique. The proposed framework helped in fusing data conflicts and redundancy over LOD Datasets. In the future we will work on new phonetic technique that takes in consideration the pronunciation and the punctuation of names which will increase the precision and recall rate and give much better results in cross language matching.

REFERENCES

- [1] J. Zhu, "Duplicate Record Detection," in *Elsevier*, 2012.
- [2] T. H. C. B. Tim Berners-Lee, "Linked Data - The Story So Far," 2009.
- [3] F. N. Jens Bleiholder, "Data Fusion," in *ACM*, 2008.
- [4] T. T. Jr, "dppedia," Wikipedia, 2015. [Online]. Available: <http://dbpedia.org/>.
- [5] S. C. V. G. Rohit Ananthakrishna, "Eliminating Fuzzy Duplicates in Data," in *ACM*, 2002.
- [6] Berners-Lee, "Semantic Web Road Map," in *W3 organization*, 1998.
- [7] L. & A.-K. Saleh, "AraTation: An Arabic Semantic Annotation Tool," 2009.
- [8] A. R. A. R. I. Majdi Beseiso, "A Survey of Arabic Language Support in Semantic Web," in *International Journal of Computer Application*, 2010.
- [9] M. H. P. Cheatham, "The role of string similarity metrics in ontology alignment," in *Tech. rep., Kno.e.sis Center*, 2013.
- [10] D. V. M. a. L. D. Dey, "Efficient Techniques for Online Record Linkage," in *IEEE Transactions on Knowledge and Data Engineering*, IEEE, 2011, pp. 373-387.
- [11] P. H. A. P. S. K. V. a. P. Z. Y. Prateek Jain, "Ontology Alignment for Linked Open Data," in *Springer*, 2010.
- [12] D. S. L. C. a. Dr. Andrew T. Freeman, "algorithm, Cross linguistic name matching in English and Arabic: a one to many mapping extension of the Levenshtein edit distance," in *ACM*, 2006.
- [13] M. C. a. P. Hitzler, "String Similarity Metrics for Ontology Alignment," in *Kno.e.sis Center, Wright State University, USA*, 2014.
- [14] P. G. I. V. S. V. Ahmed K. Elmagarmid, "Duplicate Record Detection," in *IEEE*, 2007.

- [15] R. W. P. L. K. F. W. Ho Chung Wu, "Interpreting TF-IDF term weights as making relevance decisions," *ACM Transactions on Information Systems*, vol. 26, no. 3, June 2008, p. 13, 2008 .
- [16] P. V. d. V. J. D. P. R. L. M. Peter F. Browen, "Class-Based n-gram Models of natural Language," in *ACM*, 1992.
- [17] P. K. ., C. G. a. J. N. Wilbert Heeringa, "Evaluation of string distance algorithms for dialectology," in *Linguistic Distances*, Sydney, Association for Computational Linguistics, 2006, pp. 51-62.
- [18] M. V. V. a. A. E. Elfeky, "TAILOR: a record linkage toolbox," in *Proceedings of the 18th International Conference on in Data Engineering*, 202.
- [19] W. Yancey, "Bigmatch: A Program for Extracting Probable Matches from a Large File for Record Linkage," Bureau of the Census, US, 2002.
- [20] P. Christen, "Febrl: a freely available record linkage system with a graphical user interfac," *the second Australasian workshop on Health data and knowledge management* , vol. 80, no. Australian Computer Society, pp. 14-25, 2008.
- [21] P. T. C. a. M. H. Christen, "Febrl – A Parallel Open Source Data Linkage System," in *Advances in Knowledge Discovery and Data Mining*, Berlin, Springer, 2004, pp. 638-647.
- [22] A. H. Y. A. H. Tarek El Tobely, "Web-based Arabic/English Duplicate Record Detection with Nested Blocking Technique," in *IEEE*, Cairo, 2014.
- [23] datahub, "Yago," CKAN , 2013. [Online]. Available: <http://datahub.io/dataset/yago>.
- [24] S. e. a. Zaidi, "A Cross-language Information Retrieval: Based on an Arabic Ontology in the Legal Domain," in *International Journal of Computer Applications*, 2009.
- [25] Russel, "Soundex". USA Patent 1261167, April 1918.
- [26] S. B. E. J. D. G. a. O. F. Syed Uzair Aqeel, "On the Development of Name Search Techniques for Arabic," in *WILEY interScience*, Chicago, 2006.
- [27] L. M. Dan Brickley, "FOAF Project," FOAF, 2014. [Online]. Available: <http://www.foaf-project.org/>.
- [28] R. G. B. M. D Brickley, RDF Vocabulary Description Language 1.0: RDF Schema, W3C, 2004.
- [29] F. v. H. Deborah L. McGuinness, owl web ontology language overview, W3C, 2004.

Exploring Domain Interrelations in Freebase Schema Using Modularity-Based Community Detection

Mahmoud Elbattah

College of Engineering and Informatics, National University of Ireland
m.elbattah1@nuigalway.ie

Mohamed Roshdy

Faculty of Computer and Information Sciences Ain Shams University, Cairo, Egypt
mroushdy@cis.asu.edu.eg

Mostafa Aref

Faculty of Computer and Information Sciences Ain Shams University, Cairo, Egypt
aref_99@yahoo.com

Abdel-Badeh Salem

Faculty of Computer and Information Sciences Ain Shams University, Cairo, Egypt
abmsalem@yahoo.com

Abstract— Freebase is intended to be an important component of the Linked Open Data (LOD). The paper presents a graph-driven methodology for the analysis and visualisation of Freebase complex schema. First, the methodology utilises Freebase schema types, “Included Type” relationships and “Instance Count” properties to construct a directed weighted graph schema. Second, the schema graph is employed to conduct modularity-based analysis in order to detect communities underlying Freebase schema. In view of that, the detected communities are effectively used for the purpose of revealing unobserved or implicit domain relationships.

Keywords—Linked Open Data; Community Detection; Freebase

I. INTRODUCTION

Freebase is a large, collaboratively database of cross-linked data developed by Metaweb Technologies [1]. Freebase has incorporated the contents of several large, openly accessible data sources, such as Wikipedia and Musicbrainz, allowing users to add data and build structure by adding metadata tags that categorise or connect items.

On the other hand, the massive amount of Freebase data raises an inevitable demand for effective data analysis and visualisation. Unlike other significant endeavours for exploring and visualising Freebase data such as “Thinkbase” [2] [3] and “GraphCharter” [14], the paper focused solely on Freebase schema. The paper adopted a graph-driven approach for representing the complex schema of Freebase. Furthermore, modularity-based analysis was utilised in order to detect communities in Freebase schema. The detected communities are used to explore the interrelations among Freebase domains. Specifically, we claim the following contributions:

- Utilising community detection in order to reveal unobserved or implicit domain interrelationships in Freebase schema, which has not been addressed before, to the authors' best knowledge.
- Exploring the densely connected domain communities in Freebase schema, based on the “Included Type” relationships.
- Identifying the highly interrelated domains of Freebase schema that tend to be located in numerous communities.
- Furthermore, the study provides methodological lessons concerning constructing Freebase schema as directed weighted graph using “Included Type” relationships and “Instance Count” property.

II. METHODOLOGY

A. Representation of Freebase Schema as Directed Weighted Graph

The Freebase schema was constructed as a graph, where the graph is broken down into the following components:

1. Nodes: Each node in the schema graph represented a Freebase type. The total number of graph nodes reached 1,659.
2. Edges: Linking nodes with directed edges was realised by using the “Included-Type” relationships. For instance, since the “Author” type included the “Person” type, therefore a directed edge was constructed denoting “Author” as the source node, and “Person” as the destination node. The total number of directed edges was 2,837. Figure (1) depicts an example of the included-type relationship.

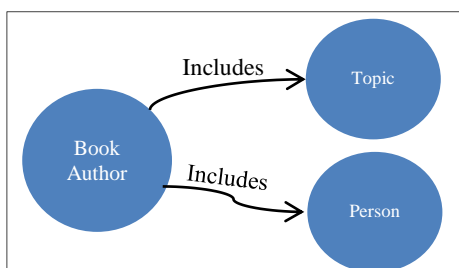


Fig. 1. An Example of how nodes were linked in the schema graph through directed edges that represent the included-type relationships.

3. Assignment of Edge Weights: The edge weight was used to indicate the relative influence of a source type on its included type. The “Instance Count”, a schema property of Freebase, was considered for that purpose. Specifically, the edge weight is represented as the ratio of the source type instance count to the included type instance count. The edge weight is defined in equation (1) as follows:

$$W = (IC_{\text{Source Type}} / IC_{\text{Included Type}}) \quad (1)$$

Where

$W \rightarrow$ Edge Weight

$IC_{\text{Source Type}} \rightarrow$ Instance Count of the Source Type
(Source Node)

$IC_{\text{Included Type}} \rightarrow$ Instance Count of the Included Type
(Destination Node)

B. Visualisation of the Schema Graph

The constructed schema graph was utilised for the purpose of visualisation. Figure (2) illustrates the schema graph with emphasis on the highest degree nodes. The graph analysis and visualisations were conducted using Gephi [5].

Gephi is an open-source software for network exploration and manipulation. According to [14], Gephi modules can

import, visualise, spatialise, filter, manipulate and export all types of networks. The visualization module uses a special 3D render engine to render graphs in real-time, using the computer graphic card. It can deal with large networks (i.e. over 20,000 nodes), because it was built on a multi-task model taking advantage of multi-core processors.

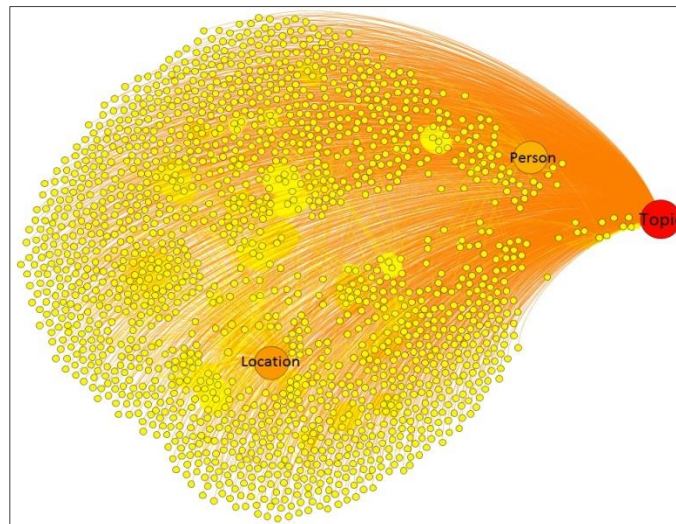


Fig. 2. Freebase schema graph with emphasis on significantly high ranked in-degree nodes. The rank of the node in-degree is represented as the node background colour ranging from yellow (lower in-degree) to red (higher in-degree). The edge directions are highlighted by the colour of source nodes.

C. Minimisation of the Schema Graph

The Schema graph was refined to present a higher view of the schema objects relationships, which is domain-based. The domain-based schema could provide a less complex graph providing an elevated perspective of Freebase schema objects interrelations. Moreover, the significantly lower number of Freebase domains (82) compared to that of Freebase types (1,659) directly contributed to decrease the complexity of the problem, and the following graph-based analysis.

For the purpose of schema minimisation, a new property needed to be added to Freebase schema, which is “Collective Instance Count”. The collective instance counts were used to assign weights to edges of the minimised graph. Collective instance count accumulatively summed the instance counts of all types associated with a specific domain. For instance, the collective instance count of “Film” domain approximately reached 4,700,00 by adding up all the instance counts of the underlying types such as “Film director”, “Film actor”, “Film producer”. The edge weight is defined in equation (2) as follows:

$$W = (CID_{\text{Source Domain}} / CID_{\text{Included Domain}}) \quad (2)$$

Where

$W \rightarrow$ Edge Weight

CID_{Source Domain} → Collective Instance Count of the Source Domain (Source Node)
 CID_{Included Domain} → Collective Instance Count of the Included Domain (Destination Node)

As a result, the number of schema graph nodes decreased from 1,659 to 82. More importantly, the number of directed edges was reduced approximately by 90% from 2,837 to 274.

D. Visualisation of the Minimised Schema Graph

The minimised schema graph was re-visualised with respect to the domain-based perspective, as shown in figure (3). The graph nodes represent Freebase domains, and the directed edges represent the included-type relationships. In addition, figure (4) demonstrates the top 10 ranked Freebase domains by the in-degree measure.

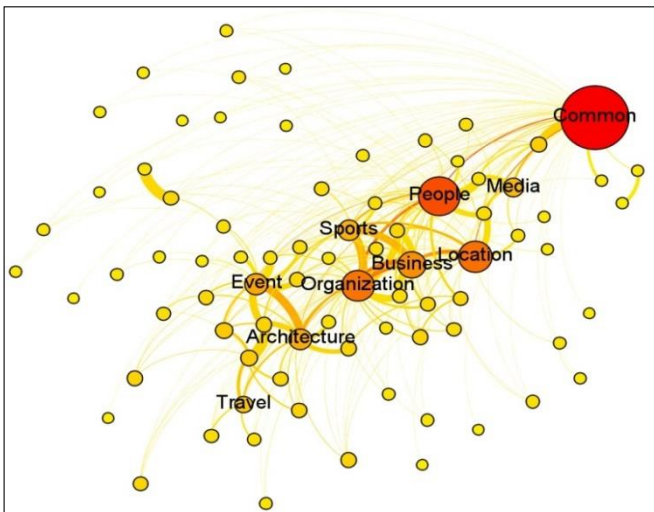


Fig. 3. Domain-based visualisation of Freebase schema graph, with emphasis on significantly high in-degree nodes. The rank of the node in-degree is represented as the node background colour ranging from yellow (lower in-degree) to red (higher in-degree). The edge directions are highlighted by the colour of source nodes.

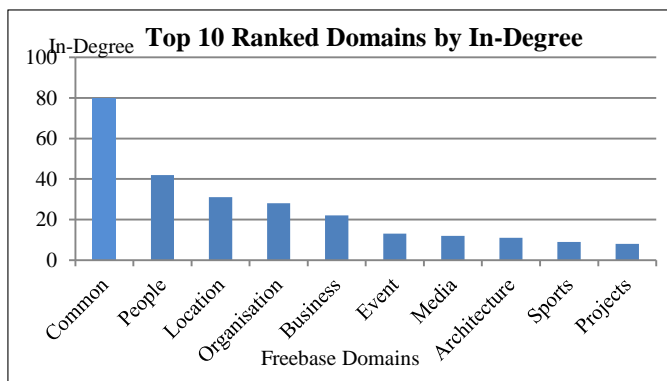


Fig. 4. Top 10 ranked Freebase domains by node in-degree. The “Common” domain has the significantly highest in-degree.

E. Normalising the Impact of High-Degree Nodes

The modularity-based analysis was adopted for detecting potential communities in the schema graph. However, the measure of modularity [6] is based on a principle that the connectivity within a community should be high, and the connectivity among communities should be low. Therefore, the negative impact of high-degree nodes should be normalised first before conducting the modularity analysis. The need for removing the higher degree nodes was acknowledged in a similar study [4] for summarizing large-scale database schemas using community detection as well.

Accordingly, the highest degree node was excluded from the schema graph, which represented the “Common” domain. As a result, the number of graph nodes and edges were reduced once again. The number of nodes decreased to 71, the exclusion of the “Common” domain resulted in the omission of other domains that had exclusive links to “Common”. Eventually, the number of edges was reduced to 197.

F. Modularity-Based Analysis

The paper adopted the algorithm presented in study [7] for conducting the community detection, which was based on modularity measure. The selected algorithm was applied in different studies related to complex network analysis such as [10], [11], [12] and [13]. The modularity measure of weighted networks, which applies to the constructed Freebase schema graph, is defined in equation (3) according to [8]:

$$Q = \frac{1}{2m} \sum_{i,j} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j) \tag{3}$$

Where

- A_{ij} → The weight of the edge between i and j
- $k_i = \sum_j A_{ij}$ → The sum of edge weights attached to vertex i
- c_i → The community to which vertex i is assigned
- $\delta(u, v)$ is 1 if $u = v$ and 0 otherwise, and $m = (\frac{1}{2}) \sum_{ij} A_{ij}$

The modularity-based analysis detected five densely connected communities. Figure (5) illustrates the five detected communities. Table (1) summarises the detected communities, the count of domains associated with each community and the included domains.

TABLE.1 SUMMARY OF DETECTED COMMUNITIES.

Detected Community #	No. of Included Domains	Included Domains
1	26	Architecture, Travel, Amusement Parks, Zoos and Aquariums, Fashion; Clothing and Textiles, Military, American football, Olympics, Tennis, Skiing, Cricket, Event, Time, Aviation, Transportation, Spaceflight, Automotive, Projects, Theatre, Opera, Books, Law, Religion, Conferences and Conventions, Royalty and Nobility, Engineering
2	21	Education, Organization, Government, Language, Business, Digicams, Food & Drink, Soccer, Sports, Ice Hockey, Baseball, Basketball, Medicine, Computers, Meteorology, Biology, Astronomy, Location, Protected Places, Rail, Bicycles
3	17	Media, Film, Music, TV, Physical Geography, Visual Art, Video Games, Fictional Universes, Internet, Comics, Games, Awards, Hobbies and Interests, Geology, Periodicals, Comedy, People
4	2	Martial Arts, Boxing
5	2	Broadcast, Radio

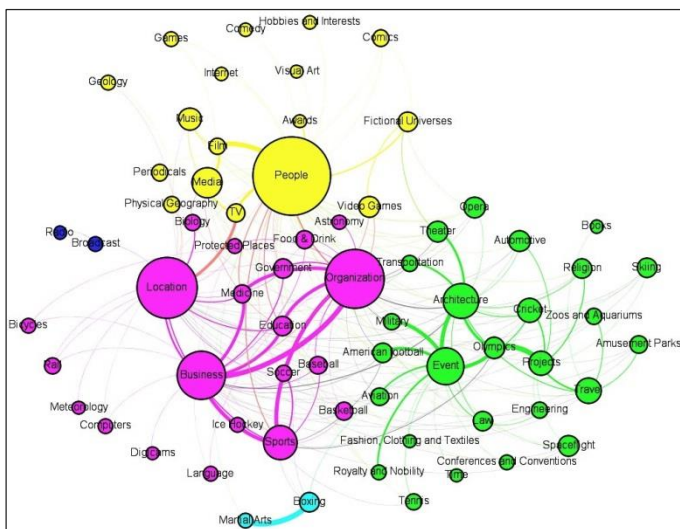


Fig. 5. Detected communities according to the modularity analysis. Each community is assigned a different colour for the purpose of demonstration.

G. Measuring Similarity Between Detected Communities and Freebase Categories

Freebase schema already includes a particular object as a grouping of related domains, which is “Category”. The Freebase categories were considered as explicit communities to be compared with the implicit (detected) communities. However, the domains underlying Freebase categories could not be found explicitly neither on Freebase.com nor other reference, to the authors' best knowledge. Therefore, the domains of each category had to be extracted using MQL queries, below is an example of retrieving domains in “Science & Technology” category. Additionally, table (2) demonstrates the extracted domains of Freebase categories.

MQL Example: MQL query to retrieve domains of “Science & Technology” category:

```

[[
  "id": null,
  "name": null,
  "type": "/freebase/domain_profile",
  "category": {
    "id": "/en/science_technology" }
]]
    
```

TABLE. 2 FREEBASE CATEGORIES AND INCLUDED DOMAINS.

Freebase Category Name	Included Domains
Science & Technology	Medicine, Computers, Meteorology, Biology, Spaceflight, Internet, Astronomy, Chemistry, Geology, Engineering, Physics
Arts & Entertainment	Film, Music, Books, TV, Broadcast, Visual Art, Video Games, Theatre, Opera, Fictional Universes, Comics, Media, Games, Radio, Periodicals
Sports	Soccer, American football, Basketball, Sports, Ice Hockey, Baseball, Tennis, Cricket, Martial Arts, Olympics, Skiing, Boxing
Society	Education, Government, Language, People, Organization, Law, Religion, Awards, Conferences and Conventions, Influence, Library, Exhibitions, Celebrities, Royalty and Nobility
Products & Services	Food & Drink, Business, Digicams, Automotive
Transportation	Aviation, Transportation, Spaceflight, Boats,

	Automotive, Rail, Bicycles
Time & Space	Location, Measurement Unit, Physical Geography, Time, Protected Places, Event
Special Interests	Architecture, Military, Travel, Amusement Parks, Zoos and Aquariums, Hobbies and Interests, Fashion-Clothing and Textiles, Symbols

Subsequently, the Jaccard similarity coefficient (Jaccard Index) was employed to measure the similarity between the implicitly detected communities and the explicitly defined categories by Freebase. The Jaccard index measures similarity between two finite sample sets as defined in equation (4) according to [9]:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (4)$$

III. RESULTS

The similarity measurement produced 40 Jaccard indices. Table (3) presents the values of Jaccard indices. In addition, figure (6) plots the Jaccard indices against the detected communities.

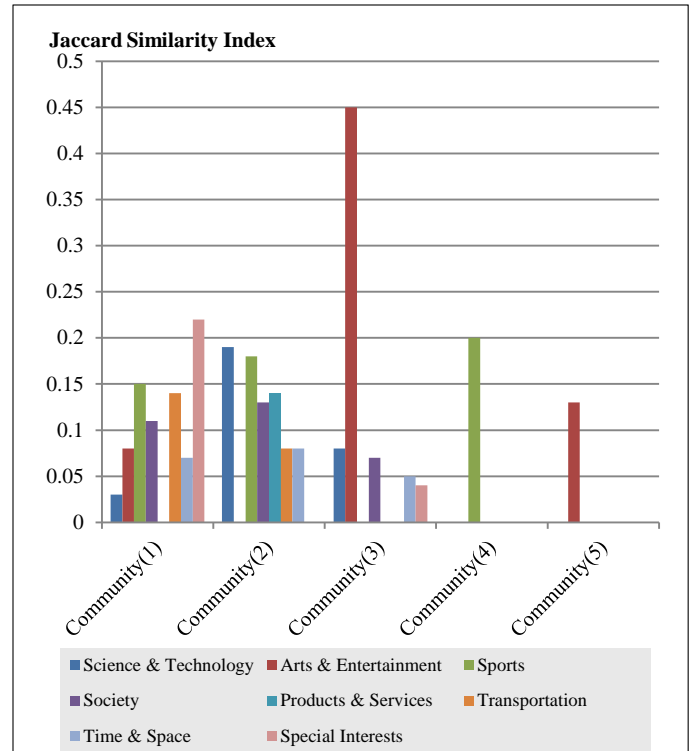
TABLE. 3 JACCARD SIMILARITY COEFFICIENTS OF DETECTED COMMUNITIES.

Community #	Jaccard Similarity Coefficients								
	Sc. & Tech.	Art & Ent.	Sport	Soc.	Prod. & Serv.	Trans.	Time & Space	Spec. Interest	
1	0.03	0.08	0.15	0.11	0	0.14	0.07	0.22	
2	0.19	0	0.18	0.13	0.14	0.08	0.08	0	
3	0.08	0.45	0	0.07	0	0	0.05	0.04	
4	0	0	0.2	0	0	0	0	0	
5	0	0.13	0	0	0	0	0	0	
	Average Similarity								
	0.06	0.13	0.11	0.06	0.03	0.04	0.04	0.05	

Fig. 6. Plotting Jaccard similarity indices against detected communities.

IV. DISCUSSION

Based on the Jaccard similarity measurements, the detected communities tended to have higher similarity with the categories of “Arts & Entertainment” and “Sports”. However, identical or relatively large similarity was not expected, which can be justified that domain implicit interrelations are not explicitly established within Freebase categories. For instance, the first community included diverse domains from different Freebase categories, which are “Society”, “Sports”, “Time & Space”, “Special Interests”, “Transportation”, “Arts & Entertainment”, “Science & Technology”. The diversity of domains included in the first community can depict the underlying interrelationships originating from the included-type relationships. However, the similarity indices can be considered as an indicator to the highly clustered communities, such as the third community.



Furthermore, the intensity of domain categories located in the detected communities could infer the interrelationships between Freebase domains. For example, the domains of “Sports” and “Society” categories can be considered to be highly involved or interrelated with other domains in diverse categories. On the contrary, “Products & Services” domains are exclusively located in an isolated community. Accordingly, the highly inter-linked domains are likely to be included in more communities. Figure (7) portrays the overlaps between the five detected communities.

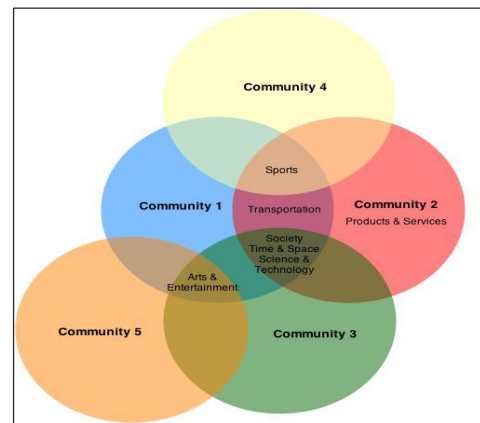


Fig.7. Overlaps between the detected communities. Highly-interrelated domains are located in intensively intersected areas, while less inter-related domains are located in fewer communities.

V. LIMITATIONS OF THE METHODOLOGY

The adopted methodology depended mainly on two particular properties of Freebase schema for constructing the schema graph, which are “Included Types” and “Instance Count”. Therefore, it might not be possible to generalise that methodology, to build other schema graphs, unless similar schema properties are available. However, the methodology can still be useful with Freebase case for the purpose of graph-based analysis or visualisation.

VI. CONCLUSIONS

In the first instance, the paper presents a graph-driven approach to analyse and visualise the large-scale schema of Freebase. The Freebase schema is represented as a directed weighted graph. Initially, the schema graph is constructed using Freebase types, included-type relationships and instance count property. Afterwards, the schema graph is minimised and restructured with respect to Freebase domains. Eventually, the impact of high-degree nodes has been normalised by excluding those nodes from the schema graph.

Secondly, modularity-based analysis is utilised to detect potential communities in Freebase schema graph. The modularity analysis could identify five densely connected communities. The Jaccard similarity indices are used to measure the similarity between the implicitly detected communities and the explicitly defined categories by Freebase. The similarity measurements can indicate that “Arts & Entertainment” and “Sports” categories have higher similarity with the detected communities. Furthermore, the overlaps between the detected communities can detect the highly inter-linked domains in Freebase schema, such as the domains of “Society” category. Hence, the community detection is demonstrated as an effective method that can reveal unobserved or implicit relationships within complex graph-based schemas, such as Freebase.

REFERENCES

- [1] Arrison, Thomas, and Scott Weidman, eds. *Steps Toward Large-Scale Data Integration in the Sciences: Summary of a Workshop*. National Academies Press, 2010.
- [2] Hirsch, Christian, John C. Grundy, and John G. Hosking. "Thinkbase: A Visual Semantic Wiki." In *International Semantic Web Conference (Posters & Demos)*. 2008.
- [3] Hirsch, Christian, John Hosking, and John Grundy. "Interactive visualization tools for exploring the semantic graph of large knowledge spaces." In *Workshop on Visual Interfaces to the Social and the Semantic Web (VISSW2009)*, vol. 443. 2009.
- [4] Wang, Xue, Xuan Zhou, and Shan Wang. "Summarizing large-scale database schema using community detection." *Journal of Computer Science and Technology* 27, no. 3 (2012): 515-526.
- [5] <https://gephi.github.io/>
- [6] Newman, Mark EJ, and Michelle Girvan. "Finding and evaluating community structure in networks." *Physical review E* 69, no. 2 (2004): 026113.

- [7] Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. "Fast unfolding of communities in large networks." *Journal of Statistical Mechanics: Theory and Experiment* 2008, no. 10 (2008): P10008.
- [8] Newman, Mark EJ. "Analysis of weighted networks." *Physical Review E* 70, no. 5 (2004): 056131.
- [9] Cesare, Silvio, and Yang Xiang. *Software similarity and classification*. Springer Science & Business Media, 2012.
- [10] Lancichinetti, Andrea, and Santo Fortunato. "Community detection algorithms: a comparative analysis." *Physical review E* 80, no. 5 (2009): 056117.
- [11] Porter, Mason A., Jukka-Pekka Onnela, and Peter J. Mucha. "Communities in networks." *Notices of the AMS* 56, no. 9 (2009): 1082-1097. Fortunato, Santo, "Community detection in graphs." *Physics Reports* 486, no. 3 (2010): 75-174.
- [12] Rubinov, Mikail, and Olaf Sporns. "Complex network measures of brain connectivity: uses and interpretations." *Neuroimage* 52, no. 3 (2010): 1059-1069.
- [13] Tu, Ying, and Han-Wei Shen. "GraphCharter: Combining browsing with query to explore large semantic graphs." In *Visualization Symposium (PacificVis)*, 2013 IEEE Pacific, pp. 49-56. IEEE, 2013.
- [14] Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy. "Gephi: an open source software for exploring and manipulating networks." *ICWSM* 8 (2009): 361-362.

The Arabic Language Status in the Jordanian Social Networking and Mobile Phone Communications

Gheith A. Abandah

Computer Engineering Department
The University of Jordan
Amman, Jordan
abandah@ju.edu.jo

Mohammed Zeki Khedher

Electrical Engineering Department
The University of Jordan
Amman, Jordan
khedher@ju.edu.jo

Waleed A. Anati

Arabic Language and Literature Department
Petra University
Amman, Jordan
anati_waleed@hotmail.com

Ahmad A. Zghoul

Department of Mathematics
The University of Jordan
Amman, Jordan
a.zghoul@ju.edu.jo

Sami M. Ababneh

Arabic Language Department
The University of Jordan
Amman, Jordan
sami_ababneh@hotmail.com

Mamoun S. Hattab

Arabic Textware
Amman, Jordan
m.hattab@arabtext.ws

Abstract—The internet and smartphone penetrations continue to rise reaching large percentages of the world populations. Likewise, many Jordanians are actively communicating through the popular social networks and mobile phone messages. There are large questions and concerns related to the characteristics and quality of the language used in these forums and how to improve it. This study addresses these issues by collecting and analyzing a large sample of text from five sources: Facebook, Twitter, news sites, blogging sites, and mobile phone short messages. We analyzed the sample comprehensively including the sender, context, message, channel, and code. We present in this paper the results related to the used language, alphabet, dialect, text components, and style. The study concludes that the bilingualism problem is manifested in Twitter and Facebook with 24% and 14% of contributions in English, respectively. Moreover, 6.4% of the analyzed Arabic samples have English words and 13.2% are written in Arabizi (Arabic in English letters and numerals). The diglossia problem is manifested as 55.4% of the sample is in colloquial Arabic, 36.4% in the standard Arabic, and 8.2% in standard with some colloquial words.

Keywords— Arabic language; Jordan; social networks; Facebook; Twitter; blogging; electronic news sites; short messages

I. INTRODUCTION

The internet, social networks, mobile phones, and smartphone penetrations are increasing year after year globally [1-3]. The Arab World and Jordan are no exceptions. More and more people are accessing the internet and social networks through their computers and smartphones. In 2014, Jordan has reached internet and mobile phone penetrations of 74% and 147%, respectively [4]. Many Jordanians are actively communicating through social networks and mobile phones. The penetration of famous social networks in Jordan such as Facebook, LinkedIn, and Twitter has reached 47.9%, 5.0%, and 2.4%, respectively [5]. In fact, Facebook is the top internet site visited in Jordan [2] and Jordanians send more than 11 million tweets monthly and have exchanged over 2.5 billion short messages last year [4]. These rates are expected to continue rising due to the rising smartphone penetration and the increasing popularity of free messaging services such as WhatsApp, Skype, and Viber.

There are also many indicators that the number and percentage of internet contributions in the Arabic language through these forums in the Arab World are increasing [5]. However, there are concerns about the quality and type of the Arabic language used in these forums and how the internet affects the language and vice versa [6].

This paper summarizes our study of the status of the Arabic language that Jordanians use in social networks and mobile phone communications. The main objectives of this study are to find the main characteristics of the Arabic language used and to identify the main problems in the quality of the language used. Hopefully, this identification would lead to solutions to improve the quality and effectiveness of Arabic language communications in these forums.

This study incorporated Jakobson's effective communication model, including the sender, context, message, channel, code, and the receiver [7]. We collected many text samples and information about their sender, context, and channel from five sources. The five sources are Facebook, Twitter, News sites comments, blogging sites, and mobile phone messaging.

There are several studies that have tackled the subjects of the Arabic language on the internet and mobile phone messaging in several Arabic countries [8-11, 17]. However,

This study handles these issues more comprehensively in Jordan by collecting large sample from five sources and analyzing this sample on many aspects as detailed below.

The details of this study are published in a long technical report [18]. This paper summarizes the methodology used in collecting fair and representative sample and analyzing this sample. Moreover, we present the analysis results related to the used language, alphabet, dialect, text components, and style.

Section II summarizes the methodology used including the developed sample collection and analysis application and the sample collection methods from the five study sources. Section III presents the results of the used language, alphabet, dialect, text components, and style. Finally, Section IV summarizes and discusses the main results, identifies three main problems, and suggests some recommendations and future work.

II. METHODOLOGY

In this section, we introduce the methodology used in this study. We describe the application developed to collect and analyze samples. We also describe how samples were collected from the five study sources.

A. Sample Collection and Analysis Application

We have developed a web-based application to facilitate and speed up the processes of sample collection and analysis. This application supports two main roles: sample *collector* and sample *analyzer*. Fig. 1 shows the main page used in sample collection. For each sample, the collector uses this page to specify the following fields.

- The sample text and topic
- The URL of the sample source webpage

Fig. 3. Language information subpage.

This paper concentrates on the analysis results of this subpage. More detail about these characteristics is in Section III.

B. Sample Collection Method

We have collected many samples from the five study sources. The collection method aimed at collecting a fair and representative sample. The following subsections describe how this sample was collected from the five study sources.

1) Facebook

Facebook is the top visited Internet site in Jordan [2]. Facebook allows its users to update their statuses, upload photos or videos, post on the walls of other users, and share and comment on almost anything posted by other users. We have collected samples of the text of the following Facebook contributions.

- Status update
- Photo or video upload description
- Posting on other's wall
- Added text of a shared contribution
- Comment on any of the above contributions

These contributions are usually related to three sources: a *user* account, a *group* of users, or a *page* of some organization, product, fans, etc. We have collected 2,507 samples of the above contributions as detailed in Table I.

TABLE I. FACEBOOK SAMPLES BY SOURCE

Facebook Source	Count	Number of Samples
User accounts	100 users	986
Groups	27 groups	752
Pages	7 pages	769
Total		2,507

The user account samples were drawn from the walls of about 100 user accounts of Jordanian users. These accounts were randomly selected using the Facebook *find friends* feature by specifying the *current city* as one of the Jordanian cities. However, this search feature is biased to the user's likely connections. To overcome this bias, we created a fresh Facebook account filled with minimal information and with no connections to get unbiased search results.

The group samples were drawn from 27 representative Jordanian Facebook groups. The interests of these groups include academic, family/tribal, cultural, political, religious, trading, sports, and hobbies.

The page samples were drawn from seven Jordanian Facebook pages that have large numbers of followers according to the lists of the *Social Bakers* site [12]. As we are interested in samples from normal users, we ignored contributions from page administrators and only collected contributions of *posts by others*.

2) Twitter

Twitter is the second most popular social networking site in Jordan [2]. Users in Twitter contribute by sending *tweets*. Each tweet is limited to 140 characters and users view the tweets of the users they *follow*. We have collected the information of 1,514 tweets using Twitter's *advanced search* feature. In order to collect fair and representative sample, we collected the samples that satisfy the following criteria.

- Original tweet, not retweet
- The twitter is a person, not an organization
- The twitter's country is Jordan
- The tweet's language is Arabic, English, or mixed

Moreover, the sample collection process extended from Jul 18, 2013 to Sep 4, 2013 over all week and day times.

3) News Sites Comments

There are more than 118 electronic press sites in Jordan [19]. Most of these sites allow the visitors to comment on the posted news. For some sites, these comments reach hundreds of comments for some popular news items.

We have collected 1,504 samples of these comments over a two-month period over all week and day times. We collected samples from various news topics including politics, economics, sports, society, arts, and culture. The sites from where these samples were collected are the sites that are most visited in Jordan [2] and allow visitor comments. The sites that we have collected samples from are Jfra News, Khaberni,

Ammon News, Alghad Newspaper, Assabeel, and Tasweer News.

4) Blogs

Blogging became popular in Jordan more than 10 years ago. Many bloggers use their blogging sites to express their views, ideas, and feelings. Many specialists think that some blogging sites such as the *Black Iris* have contributed in raising the ceiling of freedoms in Jordan [13]. However, traditional blogging is in decline as more and more bloggers are expressing themselves through Facebook and Twitter. Moreover, many Jordanian blogs are in English and reach selected segment of the Jordanian population.

Most blogging sites allow visitors to comment on the posted blogs. We have collected 52 original blogs from 52 blog sites and 459 comments on these blogs. These blogs come mainly from the most popular Jordanian blog sites according to *Jordan Blogs* and *Best Jordanian Bogs* [14, 15]. The details of these blogs are in [18].

5) Short Messages

Users of mobile phones often communicate through sending short text messages to each other. Recently, many smartphone users send such messages free of charge through specialized services such as WhatsApp, Skype, and Viber. As these messages are private from the sender to the receiver(s), we cannot collect samples of them through some open source venue. Therefore, we invited volunteers to give us samples they received on their mobile phones. To improve the fairness and representativeness of these samples, we asked each volunteer to submit 5-25 message samples that satisfy the following criteria.

- Arabic message or mixed (Arabic and English)
- Randomly selected without restriction on the message topic
- Not from some organization or some advertisement, but must be from a person

We have collected 2,502 from 141 volunteers most of them are male and female students from Jordanian universities. However, as the volunteers provided the messages they have received (not sent), the sample represents a larger segment of the Jordanian population.

III. RESULTS

In this section, we present the language information analysis of this study. The following subsections present the results found about the used language, alphabet, dialect, components, and style on the five study sources. We also comment on these results and provide some explanations.

A. Language

The language of the text Jordanians use in the five study sources is Arabic, English, or mixed Arabic and English. This study concentrated on the samples that use either Arabic or mixed language. However, we have counted the number of samples encountered in this study that use pure English. In

Facebook and Twitter, 14% and 24% of the users' contributions are in English, respectively.

Fig. 4 shows the distribution of samples that are not in English. The figure also shows the average of the distribution of the five sources. More than 95% of these samples are in Arabic and less than 5% are in Arabic with some English words or phrases. The lowest percentages of mixed language are observed in the News and Blogs (0.8% and 2.8%) and the largest percentages are in Twitter and Messages (7.3% and 6.4%).

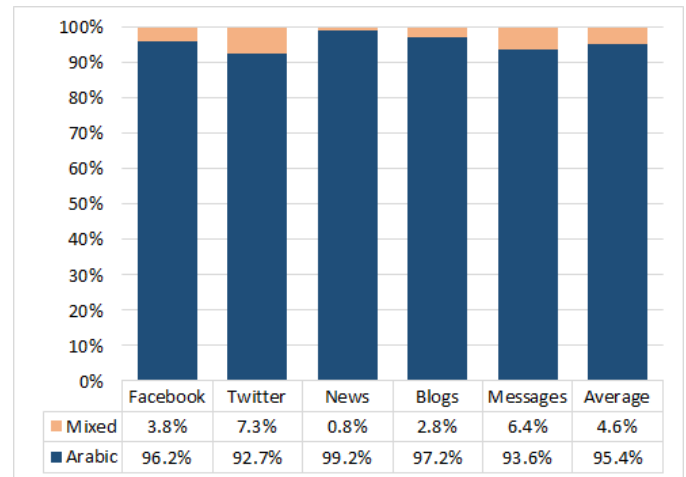


Fig. 4. Language used: Arabic or Arabic with some English words.

We explain the differences in using the mixed language among the five sources by the following points.

- The generally formal communication using news comments and blogs involves better attention to the language and using fewer foreign words.
- As tweets are limited to 140 characters each, the users strive to express their ideas with minimal characters and often use special Twitter features such as @name to draw the attention of some user and #keyword to hash tag their post with the intended subject's keyword.
- The entry difficulties in mobile phones (explained further in the next subsection) result in using higher percentage of English words in Messages.

B. Alphabet

Arabic is usually written using its alphabet that has 28 basic letters [20]. However, due to technical issues, some writers write Arabic using English letters. Currently, many people write Arabic using English letters and numerals. This writing style is called *Arabizi* [16]. Basically, The Arabic letters that have English counterparts are written using their English counterparts, e.g., 's' for Arabic **Seen** (س) and 'b' for **Beh** (ب). The rest Arabic letters are written using English letter combinations, e.g., 'th' for **Thal** (ث) and 'sh' for **Sheen** (ش), or using numerals that are closest to them in shape, e.g., '3' for **Ain** (ع) and '7' for **Hah** (ح).

This mixed writing allows the users of Twitter, Facebook, and Messages to express their feelings efficiently with small number of characters. For example, instead of writing “I feel happy” one can enter :).

As the figure suggests, this mixed text is more common in personal communication. It complements the text messages directed to others to include feelings that the body language usually reflects in face-to-face communication. Therefore, this mixed text is less common in the formal communication of News and Blogs.

E. Language Style

Fig. 8 shows the distributions of the samples according to the rhetorical style used. On average, 82.6% of the samples use the normal or plain style. This is expected as most of these samples are related to direct day-to-day communications. However, Twitter users that are usually educated and pay good attention to their tweets often use the metaphor and cynical styles at 17.5% and 17.1%, respectively.

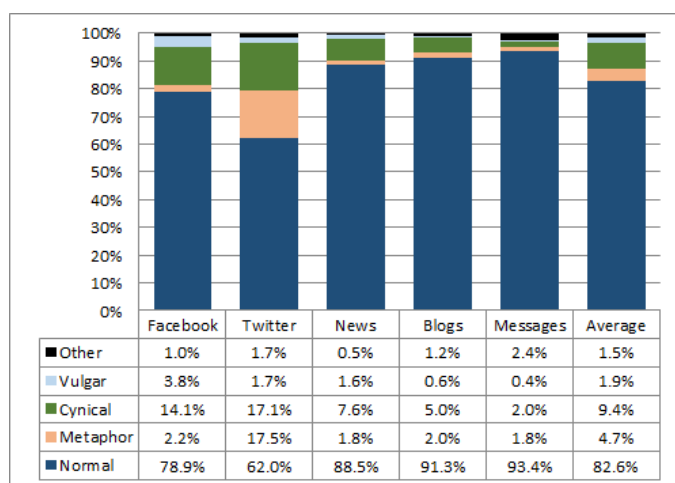


Fig. 8. Language style: normal, metaphor, cynical, vulgar, or other.

The cynical style is also common in Facebook and News at 14.1% and 7.6%, respectively. We think that this is an interesting phenomenon worth of further investigation. We wonder how this phenomenon is related to the hard Jordanian political, economic, and social situations.

Finally, the percentage of the vulgar style is low at an average of 1.9%. This indicates that most contributions observe politeness and good manners. However, Facebook has the highest percentage of this style at 3.8%.

IV. DISCUSSION AND CONCLUSIONS

In this paper, we have described the methodology used in a comprehensive study of the status of the Arabic language in the social networking forums and mobile phone messaging. This study collected a large sample of 8,538 Jordanians' contributions in the five study sources: Facebook, Twitter, News, Blogs, and Messages. These samples were analyzed taking into consideration Jakobson's communication model.

This paper presents the study results related to the code and some message aspects, specifically the results of the language information analysis. We can summarize the results for each source as follows.

- **Facebook** has significant percentages of messages in Arabizi and colloquial Arabic at 8.2% and 67.7%, respectively. Users often include symbols in their text (22.1%) and mainly use the normal and cynical styles (78.9% and 14.1%).
- **Twitter** has about 24% of the tweets in English and 7.3% of the remaining tweets are in Arabic with English words. Arabizi and colloquial percentages are also significant at 7.9% and 50.1%, respectively. Twitter users show the highest use of symbols (37.3%) and the highest use the metaphor and cynical styles (17.5% and 17.1%).
- **News** has the lowest percentages of message in Arabic with English words, using Arabizi, and with symbols at 0.8%, 0.3%, and 6.8%, respectively. However, it has highest percentages of messages in standard Arabic and standard Arabic with some colloquial words at 67.2% and 17.9%, respectively.
- **Blogs** also has low percentages of message in Arabic with English words, using Arabizi, and with symbols at 2.8%, 4.4%, and 15.3%, respectively. And high percentages of messages in standard Arabic and standard Arabic with colloquial words at 56.9% and 13.3%, respectively.
- **Messages** has the highest percentages of messages using Arabizi, in colloquial Arabic, and in the normal style at 31.0%, 75.9%, and 93.4%, respectively. The use of symbols is also high at 19.2%.

These results show that there are the following three problems related to the status of the Arabic language on these forums.

- **Bilingualism Problem:** In addition to Arabic, English is highly present in these forums. The contributions of Jordanians in Twitter and Facebook are 24% and 14% in English, respectively. Moreover, the analyzed Arabic contributions from the five sources show that 6.4% of the messages have English words and 13.2% are in Arabizi.
- **Diglossia Problem:** The colloquial Arabic is common in conversation and casual communications. The standard Arabic, on the other hand, is used in formal communications. These dual dialects were observed in the five study sources at averages of 55.4% in colloquial Arabic, 36.4% in standard Arabic, and 8.2% in standard with colloquial Arabic.
- **Linguistic Weakness Problem:** This problem is not presented in this paper, but the study shows that there is high percentage of contributions that have weak Arabic language. This weakness is manifested in large

rates of spelling, lexical, morphological, and grammatical errors.

We think that these problems can be mitigated by technical and nontechnical solutions, including legislative, informational, and educational solutions. The technical solutions should concentrate on improving how Arabic text is efficiently entered, especially on mobile phones. Moreover, there is a great need to improve the operating systems and applications' support of the Arabic language. Arabic spell and grammar checkers, for example, are not available or expensive. Developing and freely providing such support would definitely improve the Arabic language on such communication forums.

Finally, and as ideas for future work, we are interested in studying these issues in other Arabic countries and even for Arab communities in foreign countries. Moreover, we are interested in monitoring these issues over time and studying the effect of technological advancements in smartphones and communications on these issues.

ACKNOWLEDGMENT

This research was organized and financed by the Jordan Academy of Arabic and the Jordanian National Committee for the Advancement of the Arabic Language towards the Knowledge Society. The authors are grateful for the financial support of this research.

REFERENCES

- [1] Internet Live Stats, Internet Users by Country (2014), <http://www.internetlivestats.com/internet-users-by-country/>, Jul 1, 2014.
- [2] Alexa, Top Sites in Jordan, <http://www.alexa.com/topsites/countries/JO>, last visited Apr 23, 2014.
- [3] M. Kakihara, "Grasping a global view of smartphone diffusion: An analysis from a global smartphone study," Int'l Conf. on Mobile Business, London, 2014.
- [4] Telecommunications Regulatory Commission of Jordan, Telecommunications Indicators: 2014/Q3, <http://www.trc.gov.jo>, last visited Jan 22, 2015.
- [5] Arab Social Media Report, Citizen Engagement and Public Services in the Arab World: The Potential of Social Media, 6th ed., The Governance and Innovation Program, Mohammed Bin Rashid School of Government, Dubai, available on: www.ArabSocialMediaReport.com, June 2014.
- [6] D. Crystal, Language and the Internet. Cambridge, UK, Cambridge University Press, 2001.
- [7] R. Jakobson, "Closing statement: Linguistics and poetics," in Style in Language, T. A. Schocok, Ed. New York: Wiley, 1960, pp. 350-373.
- [8] M. Wrschauer, G.R. ELSaid, and A. Zohry, "Language choice online: Globalization and identity in Egypt," J. of Computer-Mediated Communication, vol. 7, no. 4, Jul 2002.
- [9] R.A. Abdulla, "Arabic language: Use and content on the internet," Bibliotheca Alexandrina Access to Knowledge Toolkit I, pp.124-140, 2009.
- [10] R. Peel, "The internet and language use: A case study in the United Arab Emirates," Int'l J. on Multicultural Societies, vol. 6, no. 1, pp. 146-158, 2004.
- [11] M.A. Al-Khatib and E.H. Sabbah, "Language choice in mobile text messages among Jordanian university students," SKY J. of Linguistics, vol. 21, pp. 37-65, 2008.
- [12] Social Bakers, Jordan Facebook Page Statistics, <http://www.socialbakers.com/facebook-statistics/jordan>, last visited Jan 15, 2015.

- [13] The Black Iris of Jordan, <http://black-iris.com>, last visited Jan 18, 2015.
- [14] Araboo, Jordan Blogs, <http://www.araboo.com/dir/jordan-blogs>, last visited Apr 23, 2014.
- [15] Squidoo, Best Jordanian Blogs, <http://www.squidoo.com/jordan-blogs>, last visited Apr 23, 2014.
- [16] M. A. Yaghan, "Arabizi: A contemporary style of Arabic slang," Design Issues, vol. 24, no. 2, pp. 39-52, 2008.

REFERENCES IN ARABIC

- [17] و. المنصور، "من استعمالات اللغة المحدث (العربيزي)،" وقائع مؤتمر اللغة العربية ومواكبة العصر، الجامعة الإسلامية بالمدينة المنورة، 2013.
- [18] اللجنة الوطنية الأردنية للتهوض باللغة العربية للتوجه نحو مجتمع المعرفة، رصد واقع اللغة العربية في ميدان التواصل على شبكة الإنترنت والهاتف المحمول - دراسة علمية ميدانية تحليلية - الواقع والمأمول، مجمع اللغة العربية، أيلول 2014.
- [19] ويكيبيديا، ملحق: قائمة الصحف الإلكترونية الأردنية، <http://ar.wikipedia.org>، زيارة 18 كانون ثاني 2015.
- [20] غ. عبندة وف. خندقجي، "قضايا في رمز الحروف العربية"، العلوم الهندسية، مجلة دراسات، مجلد 31، عدد 1، ص. 165-177، 2004.

Similar Searching, Unsimilar Clicking

Nikolai Buzikashvili

Institute of System Analysis
Russian Academy of Sciences
Moscow, Russia
buzik@cs.isa.ru

Abstract— In the study, a search engine log is partitioned into IP classes that differently present audiences of free and busy searchers. It is shown that searching behavior of users from different classes is practically identical in all characteristics except their click behavior. Differences in click behavior between classes are great. Free users click more frequently than busy users while search in the same manner on other counts.

Keywords— *click behavior; IP address; query log analysis; query reformulation*

I. INTRODUCTION

A lot of studies look at click behavior and its uniform models. Some studies research individual and task differences in click behavior of different users ([3], [4], [5], [9]). Web log analysis reveals different manners of real-life search and clicking, but the effect of individual/task differences and environmental factors on these differences is not clear from this analysis.

Psychological studies (e.g., [8]) investigate the effects of work environment on the worker's performance. Several papers investigate web search behavior depending on different environmental conditions (e.g., [6], [7]).

In this web log-based study, we indirectly reveal the influence of real-life environmental factors and compare search behavior of users who are mainly "stressed-out office workers" or mainly "free homebodies". We can assume that when an office worker search is aimed at his office responsibilities rather than personal goals he must perform other functions and may be overloaded by them. We presume that his searching behavior under office stressors differs from his behavior in a relaxed atmosphere. We also presume that the shift of search behavior under varying environmental conditions reflects not so much individual manners but a common environmental dependency.

We investigate not the influence of individual features such as "advanced user" or "newbie" but the influence of the real-life environmental factors on search behavior of [the same] searchers.

The opportunity of this long-scale real-life comparison is provided by the "IP classes" consideration. The IP classes of users are defined in the following way: the IP- N class includes all users who share the same IP address with ($N-1$) other users. There are two reasons why a search engine detects different

users operating from the same IP: (1) different browsers on the same PC, (2) different PCs sharing the same proxy server. As a result, IP-1 corresponds to individual users, IP-2 may correspond either to two users who use different browsers on the same PC or to a proxy presenting two PCs, and the IP-3+ classes mainly correspond to users sharing the same proxy.

A fraction of (from-)home users among the IP-1 users appears to be bigger than among the IP-3+ (i.e. proxy) users. Of course, there is no strong relation between home/office and IP-1/IP-3+. Some of the IP-99 users may be cliff dwellers whilst an IP-1 user may be a small enterprise worker. However, the rate of cookies created on weekend (see Table 1) is a good indirect indicator of fractions of home/office users, and this rate in the IP-1 class is bigger than in IP-3+ classes.

Also, while there is no a strong relation between an office and noisy environment, and home and relaxed atmosphere (the counterexamples are obvious just as some fishes can fly and some birds cannot fly), we can assume that on average there is more stressed-out search behavior at work and more free behavior at home [8].

In the study, we consider the characteristics of search behavior of different IP classes and discover that all of them are very similar except that of the click activity. Frequencies of clicks per query or per first page of the retrieved results are about 30% bigger for the IP-1 users than for the IP-4+ users. This difference can be explained neither by the topical difference between "home" and "office" queries nor by different ways of moving across the pages of the retrieved results. Thus, the study discovers a strong non-individual difference in click behavior between "free home" users and "busy office" ones. However, we do not know whether free home squirrels/ birds visit useless sites or busy office workers ignore the necessary ones.

II. IP CLASSES

Web users may share the same IP address either if they use different browsers on the same PC or they operate via the same proxy server. If 3+ users have the same IP address, they probably use a common proxy rather than 3+ browsers on the same PC.

IP group. An IP group is a set of users who submit queries from the same IP address.

IP class. If an IP group includes N users, the users are attributed to the IPN class. We also refer to queries, sessions, clicks, etc. of users belonging to the IPN class as IPN class queries, sessions, etc

III. RESEARCH QUESTIONS AND ASSUMPTIONS

Research Question

Are real-life working conditions stressful enough to change user search behavior? In particular, are the following features different in (differently stressed) IP classes:

- *A query level:*
 - terms per a query,
 - clicks per a query,
 - viewed pages (screens) of the retrieved results per a query,
 - clicks per viewed pages,
 - clicks on the first page of the results
- *A search session level:*
 - task sessions per a temporal session,
 - queries per a task session,
 - click behavior in different types of task sessions:
 - single-query sessions,
 - sessions with linear query modification,
 - sessions with branching query modification
 - clicks in query narrowing and broadening modifications.

Besides, if the clicks and viewed pages of the retrieved results are actually different in the “office” IP-3, 4+ classes from the “home” IP-1 class then we should test two hypotheses “*once started they will continue*”: if a user from the “office class” made the first click then he will continue clicking similarly to a user from the “home” IP-1 class; if a user from the “office class” moved to the next page of the retrieved results he will continue moving through pages just as the user from the “home” class.

In the study of the environment influence on search, we do not focus on “a user as a sequence of his operations during a long observation period”. “User” is an irrelevant unit, not in the sense that different people can operate from the same UID but meaning that the same person may be busy or free. Even

office work has breaks and pauses and even stay-at-homers may be unexpectedly busy. Whereas *user* is an irrelevant unit, *query* is a minimum unit and *task session* (and even *temporal session*) is a convenient unit representing stable (“busy” or “free”) behavior.

Assumption

We assume that:

- a fraction of home users in the IP-1 class is significantly bigger than in the IP-4+ class, and a fraction of home users in the IPK class is not smaller than in the IP-($K+1$) class,
- home users are less stressed-out by their environment.

At weekends and on holiday days the degree of activity of office workers decreases more than of home users, and it can be assumed that the rate of office cookies created at weekends is less than that of home users’ cookies. The data in Table 1 support the first assumption: the rate of cookies created on Sunday decreases over the IP classes, and is twice as little for the IP-4+ class than for the IP-1 class.

TABLE I. FRACTIONS OF COOKIES CREATED ON SUNDAY

IP-1	IP-2	IP-3	IP-4+
8.5%	8.5%	7.5%	4.4%

IV. DATASETS

We use a complete one-day dataset (March 20, 2007) drawn from the logs of the *Yandex* search engine. The dataset combines three logs (a query log, a log of the results and a click-through log) and reports queries, retrieved results and clicks on them.

The users in the dataset are represented by unique UIDs where UID is a concatenation of $\langle hash(IP\ address) \rangle$ and 10-digit $\langle time_of_cookie_creation \rangle$, which allows us to detect all users who share the same IP address (the later *Yandex* public datasets formats do not provide this useful feature). According to UIDs the dataset was separated into IP-1, IP-2, IP-3, IP-4+ sets presenting corresponding IP classes.

The users who submitted more than 40 unique queries per 1-hour are eliminated as “robots”. To segment a logged time series of transactions into temporal sessions a 30 min cut-off was used. Table 2 shows cleaned-up datasets of IP classes.

TABLE II. GENERAL CHARACTERISTICS OF IP CLASS DATASETS

	IP-1	IP-2	IP-3	IP-4+
Users	656557	98790	35022	27434
Temporal sessions	1105496	163021	56064	43464
Queries unique in a temporal session	2463767	367193	125074	95815
Task sessions	1632796	241374	82264	63374

V. TERMS AND METHOD

In the paper, we use the notions of:

— a *query skeleton* that include only those query terms that are nouns, names, acronyms and unknowns which may be attributed to these parts of speech, a skeleton includes neither features (adjectives,) nor actions (verbs, adverbs),

— a *query narrowing* here denotes an expansion a query by additional term(s), and a *query broadening* here denotes an exclusion of one or more terms from a query (e.g., when a user submits query <cat> after submission of <red cat> he narrows the initial query, a modification of <cat toys> into <toys> is a broadening, and a modification of <cat toys> into <cat food> is neither narrowing nor broadening),

— a *temporal session* as a sequence of the user's transactions with the search engine cut from previous and successive sessions by a 30 min time gap;

— a *task session* as all queries of a task session technically defined as a connected component of the similarity graph of queries submitted during a time session. To extract task sessions (a) a matrix of term-based pairwise similarity of all unique queries of the current temporal session is filled (two queries are defined as similar if they contain common skeleton terms), (b) a transitive closure of this similarity relation is made [1]. The method [1] covers misprints in queries.

VI. FEATURES ON ALL QUERIES

Query Level Features

The general query-level features of IP classes are reported in Table 3.

TABLE III. FEATURES OF IP CLASSES. QUERY-LEVEL

	IP-1	IP-2	IP-3	IP-4+
terms in a query	3.02	3.01	3.01	2.98
terms in a query skeleton	2.18	2.18	2.19	2.18
viewed pages per query	1.58	1.57	1.54	1.50
clicks per query	1.64	1.50	1.35	1.26
clicks per viewed page	1.04	0.96	0.88	0.84
clicks on the first page	1.25	1.15	1.05	0.99

Users from all IP classes formulate queries identically (see “terms in a query” and “terms in a query skeleton”). The number of moves between pages of the retrieved results only slightly decreases over IP classes.

On the contrary, the click-based characteristics demonstrate big differences among classes. A number of clicks per a submitted query, a number of clicks per a viewed page of the retrieved results and a number of clicks on the first page of the retrieved results monotonously decreases over IP classes.

Session Level Features

The general session-level features of IP classes are reported in Table 4. While a number of temporal sessions during a day steadily decrease over IP classes from 1.68 at IP-1 to 1.58 at IP-4+, the average numbers of task sessions in a temporal session are similar, and the average numbers of queries in a task session are practically identical. Thus, users in the IP-3+ classes start session less often but once started they behave similarly to IP-1 users.

TABLE IV. FEATURES OF IP CLASSES. SESSION-LEVEL

	IP-1	IP-2	IP-3	IP-4+
temporal sessions per user	1.68	1.65	1.60	1.58
task sessions per temporal one	1.48	1.48	1.47	1.46
queries per task session	1.51	1.52	1.52	1.51
broadening query reformulation: in linear reformulations	11.4%	11.4%	11.4%	11.4%
in branching reformulations	23.7%	23.6%	24.1%	24%
narrowing query reformulation: in linear reformulations	38.3%	38.2%	38.5%	38.3%
in branching reformulations	14.2%	14.2%	15.4%	13.6%

Fractions of broadening and narrowing query reformulations in task sessions are the same across IP classes.

Different Types of Query Reformulation

According to [2] a task session may include several branches and a complex search task may be decomposed into chains which merge in the final step. However, the most significant types are linear and branching query modification. Here we use Jaccard part-of-speech-based similarity metric to detect inter-query dependencies in a task session. For example, branching search will be detected in the sequence of 3 queries <big red cat>, <red cat>, <big cat>, whilst a linear modification is detected in the sequence <red cat>, <cats food>, <cats veterinary>.

Table 5 shows practically identical rates of 3 types of temporal sessions among temporal sessions of IP classes.

TABLE V. TYPES OF TEMPORAL SESSIONS

	IP-1	IP-2	IP-3	IP-4+
Total	1109481	163577	56225	43569
Single-query temporal sessions	52.14%	51.35%	51.84%	51.73%
Linear temporal sessions	41.27%	41.95%	41.85%	42.09%
Non-linear temporal sessions	6.59%	6.70%	6.31%	6.18%

Table 6 shows an average number of clicks per components of single-query, linear and branching task sessions. We see two sorts of differences: (1) the difference between IP-classes and (2) replicated in all IP classes differences between different structures of search and between query position in the same structure.

Clicks dependency on a query position in a search structure is near identical among IP classes, but belonging to IP class shifts click values. For example, Fig. 1 shows an average number of clicks in positions of 2- and 3-query linear chains of linear query modification. And IP classes' plots look like results of a parallel shift of the same plot.

TABLE VI. CLICKS PER CHAIN OF QUERY REFORMULATIONS IN 3 STRUCTURES OF SEARCH

Chain Length	IP-1	IP-2	IP-3	IP-4+
Single-query task session	1.63	1.53	1.40	1.40
Linear chain of query modification				
2-query	3.38	3.08	2.70	2.56
3-query	4.62	4.11	3.72	3.54
Branches (including a root query)				
First branch				
2-query	3.14	2.84	2.74	2.22
3-query	4.38	3.81	3.78	3.66
Last branch				
2-query	3.74	3.34	3.08	2.52
3-query	4.92	4.41	4.26	3.45

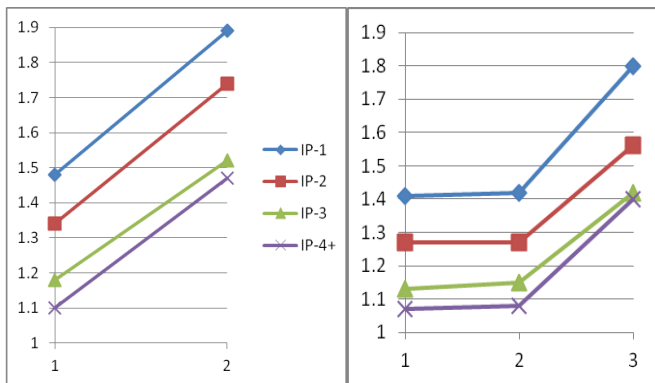


Fig. 1. Average number of clicks in positions of 2- and 3-query linear chains

Resume

No noticeable difference in search behavior is observed between classes except the only, but great difference in clicks. Any click-based query- or session-level characteristic (clicks per a query, per viewed pages, per first viewed page, clicks in any type of query modification chains) sharply decrease over IP classes.

VII. CLICKS AND VIEWED PAGES OF THE RETRIEVED RESULTS

Differences between IP Classes in Clicks and Viewed Pages Distributions.

Tables 3-6 present mean values. Let's consider clicking and moving in more detail. Distributions of clicks per query

are reported in Tables 7. Also we consider “clicks per the first page of the retrieved results” and “viewed pages per query” distributions.

We compare clicks per query distributions of the IP classes on $[L, 15+]$ intervals of clicks per query ($L=0,..,14$) by χ^2 test. A log-scaled Fig. 2 shows empirical and critical χ^2 values. A sample curve in a point L presents empirical χ^2 value for the interval $[L,15+]$, and a critical curve shows critical χ^2 value at $p=0.05$ for degrees of freedom for <4 sets, $15-L$ set size $>$. Distributions on $[L, 15+]$ are similar beginning with $L=3$.

TABLE VII. CLICKS PER QUERY DISTRIBUTIONS IN IP CLASSES

	0	1	2	3	...	15+
IP-1	1010108	650483	295077	172617	...	16671
IP-2	165656	91237	41113	23657	...	2195
IP-3	61506	29424	12994	7501	...	623
IP-4+	49190	22161	9470	5297	...	420

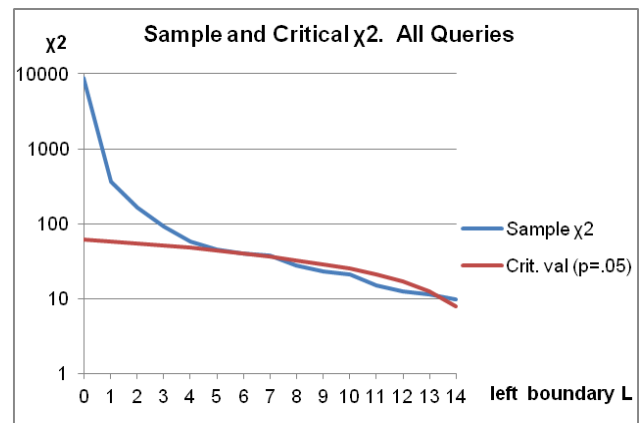


Fig. 2. Sample χ^2 ($[L,15+]$) and critical $\chi^2(p=0.05)$ as functions of the left boundary $L=0,..,14$

As Fig. 2 shows, click behaviors of IP classes strongly differ on the first clicks and are similar for subsequent clicks. Distributions of clicks on the first page of the results demonstrate just the same incoherence in the first clicks and coherence in the following clicks. At the same time, distributions of viewed pages are similar on all intervals $[L, 9+]$, where $L=1,..,8$.

Description in Terms of Transition Probabilities.

Let's return to the hypotheses “once started they will continue”. As regards to click behavior, it means that empirical probabilities of the first click may differ between classes but probabilities of transition to the any following clicks are very similar.

Empirical transition probabilities of the next click on the results of the query and transition probabilities of the next click on the first page of the results are shown in Fig. 3, where

$p(c)$ is an empirical probability of the next click when $c-1$ clicks are made.

As Fig. 3 shows, probabilities for different IP classes greatly differ for the first click and are very similar for the further clicks. The trigger “once started” hypothesis is true for click behavior. A mainly stressful environment is not totally and continuously stressful and even a normally busy worker has some free time.

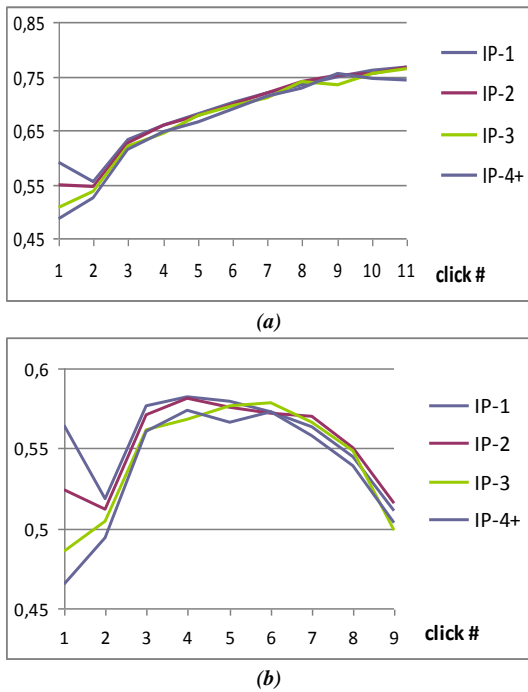


Fig. 3. Probabilities of transition to click# (a) on all viewed pages and (b) on the first page

Empirical transition probabilities of moving to the another page of the retrieved results are shown in Fig. 4, where $p(\text{move}\#)$ is an empirical probability of $\text{move}\#$ -th move to the another page (i.e. after viewing $\text{move}\#$ retrieved pages). The probabilities are similar among all IP classes and moving across the retrieved results does not depend on the IP class. This is not a surprise since viewed pages per query distributions are similar among classes.

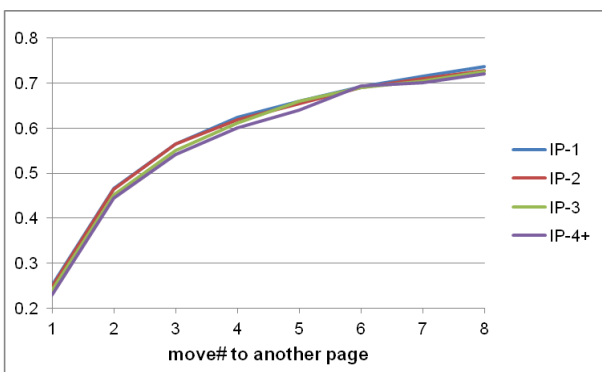


Fig. 4. Probability of $\text{move}\#$ to the another page of the retrieved results

VIII. FEATURES IN TOPIC DIMENSIONS

Different search topics lead to different search behavior. Topics may be differently presented in queries of different IP classes and between-class differences may be the result of the differences in the topic occurrence. To check a uniformity of the discovered dependencies we investigate search behavior on three topics.

We consider two topics – “Travel” and “Education”. We choose 20 topic-specific terms for the *Travel* topic and 10 topic-specific terms for *Education*. If a query contains a topic term, this query is considered as the “topical” one. If a task session contains a topical query than the session is considered as a topical session. Table 8 shows the number of topical queries, topical sessions and queries belonging to the sessions in IP classes.

TABLE VIII. TOPICAL QUERIES AND QUERIES IN TOPICAL TASK SESSIONS

	IP-1	IP-2	IP-3	IP-4+
Travel queries	45689	6246	2103	1556
Education queries	52871	7548	2434	1766
Queries in Travel sessions	68412	9399	3187	2416
Queries in Education sessions	79499	11395	3671	2624

The results in Table 9 show the same dependencies that were observed on all queries: (1) similar values of queries per task session, (2) a slightly decreasing number of viewed pages and (3) a sharply decreased number of clicks over IP classes.

TABLE IX. QUERIES IN ALL AND TOPICAL TASK SESSIONS

	IP-1	IP-2	IP-3	IP-4+
Queries per task session in:				
All sessions	1.51	1.52	1.52	1.51
Travel sessions	2.23	2.25	2.25	2.36
Education sessions	2.45	2.45	2.45	2.37
Clicks per query in:				
All sessions	1.64	1.50	1.35	1.26
Travel sessions	1.77	1.65	1.62	1.35
Education sessions	1.95	1.72	1.56	1.61
Viewed pages per query in:				
All sessions	1.58	1.57	1.54	1.50
Travel sessions	1.59	1.57	1.56	1.57
Education sessions	1.74	1.76	1.66	1.68
Clicks per viewed pages in:				
All sessions	1.04	0.96	0.88	0.84
Travel sessions	1.12	1.05	1.03	0.86
Education sessions	1.08	0.98	0.91	0.96

	IP-1	IP-2	IP-3	IP-4+
Clicks on the first page in:				
All sessions	1.25	1.15	1.05	0.99
Travel sessions	1.39	1.29	1.27	1.09
Education sessions	1.40	1.26	1.18	1.15

IX. CONCLUSION

The empirical study of IP classes gives answers to the questions about the influence of the environmental stressors on real-life search and click behavior of the Web user:

- a query formulation/reformulation does not vary across IP classes and does not depend on real-life stressors,
- the number of queries in task sessions, the number of task sessions in a temporal session and fractions of different types of task sessions do not vary across IP classes and do not depend on real-life stressors,
- the number of the viewed pages of the retrieved results decreases slightly over IP classes,
- the number of clicks (per query, per a viewed page of the retrieved results, on the first viewed page) and in all types of task sessions decreases over classes along with a fraction of free users.

The first click plays the trigger role: if a user in the “office IP class” started clicking he will continue clicking similarly to a “home” IP-1 class user (Fig. 3). 0-click behavior frequent in IP-3+ classes seems to be a bit strange: “a busy user” has no time to visit the retrieved pages but has enough time to move through 2+ pages of the retrieved results. Real-life environment stressors do not make searchers change query (re)formulation or significantly change the number of

viewed pages but urge them to decrease the number of clicks.

The rates of IP classes have changed since 2007 and the fraction of IP-4+ class must have increased. However, it is not the IP classes which matter as such but the fact that since 2007 the share of home users (retired people, home workers), i.e. the share of “free users” with their non-stressed behavioral patterns has grown.

REFERENCES

- [1] N. Buzikashvili. “Automatic Task Detection in the Web Logs and Analysis of Multitasking”, 9th Conference on Asian Digital Libraries, 2006, Kyoto, Japan, 2006, LNCS 4312, Springer-Verlag, 2006, pp. 131-140.
- [2] N. Buzikashvili. “Structure of the Web Searcher’s Query Modifications: Sequential, Branching, Merging, Re-Merging and Non-Sequential Execution”. Paper 569, 5th International Conference on Information Technology, 2011, Al Zaytoonah University, Amman, Jordan, 2011.
- [3] C. Eickhoff, J. Teevan, R. White, and S. Dumais. “Lessons from the journey: a query log analysis of within-session learning”, WSDM’14, NY, USA ACM Press, 2014, pp. 223-232.
- [4] J. Jiang, D. He, and J. Allan. “Searching, Browsing, and Clicking in a Search Session: Changes in User Behavior by Task and Over Time”, in proceedings of 37th ACM SIGIR conf. on research and development in information retrieval, Gold Coast, Australia, 2014, ACM Press, pp. 607-616.
- [5] B. Hu, Y. Zhang, W. Chen, G. Wang, and Q. Yang. “Characterizing search intent diversity into click models”, in proceedings of 20th World Wide Web Conference, Hyderabad, India. ACM Press, 2011, pp. 17-26.
- [6] S.Y. Rieh. On the Web at Home: “Information Seeking and Web Searching in the Home Environment”, J. of the Association for Information Science and Technology, vol. 55, no. 8, 2004, pp. 743-753.
- [7] S. Sushmita, H. Joho, M. Lalmas, and R. Villa. “Factors affecting click-through behavior in aggregated search interfaces”, in proceedings of 19th Conf. Information and Knowledge Management, Toronto, 2010, ACM Press, 2010, pp. 519-528.
- [8] J.C. Vischer. “The effects of the physical environment on job performance: towards a theoretical model of workspace stress”, Stress and Health, vol. 23, no 3, 2007, Wiley, pp. 175-184.
- [9] R. White, S. Dumais, and J. Teevan. “Characterizing the influence of domain expertise on web search behavior”, 2nd ACM Conf. on Web Search and Data Mining, Barcelona, Spain. ACM Press, 2009, pp. 132-141.

A Query Log-Based Study of Cross-Nation Perception

Nikolai Buzikashvili

Institute of System Analysis
Russian Academy of Sciences
Moscow, Russia
buzik@cs.isa.ru

Abstract— Query logs are a huge and solid source for sociological analysis. However, they are insufficiently used in the sociological analysis, in particular in the comparative studies of different audiences. The paper presents a study of search images of Japan in queries of Russian and U.S. Web searchers. One-day logs of the *Yandex*, the Russian search engine, and the U.S. *Excite* were automatically analyzed to detect several categories of queries referring to Japan. Users submitting Japan-referring queries were attributed to these categories. The study (a) compares rates of categorized Japan-referring users among Russian and U.S. searchers, (b) analyzes cross-linking between categories. The findings are: (a1) the Russian searchers are more interesting in Japan-referring topics, (a2) differences depend on categories: Russians show much more consumer interests, while U.S. are superior in masscult interests; (b1) the users submitting consumer queries less frequently search other topics referring to Japan; (b2) the users submitting queries relating to Japan culture more frequently search other Japan-referring topics; (b3) a Russian searcher searches several different Japan topics more frequently than U.S. searcher; (b4) the Russian and U.S. audiences significantly differ by the topic co-occurrence.

Keywords— query log analysis; cross-cultural perception

I. INTRODUCTION: SOCIOLOGY OF SEARCH

Among three questions considered by the researchers of the Web search, “Who searches the Web?” (subjects), “What do they search for?” (objects) and “How do they search?” (search tactics) the first two questions primarily relate to the applied sociology and should be formulated and answered consistently. The Web era has opened not only a new field of social activity but also a huge source of the data for sociological analysis. Query logs of Web search engines are a capacious but very special source of knowledge on public interests. Logs as such give no way to reveal either attitudes or origins of interests (except when a query is a result of another query).

While sociology of the Web mainly answers “Who searches the Web?” (age, gender, education, etc.) and uses polls, the query log-based sociology answers “What do they search for?” and uses query logs. The common subject of the Web log based sociology is a classification of queries by searched topics ([1], [2], [4], [5], [7]). More sociologically sophisticated query log based studies such as [6], [9] are so far rare.

The paper presents a comparative study based on the logs of queries submitted by two national audiences (Russian and U.S.) searching for the topics related to the third state (Japan) and its culture. In the study we use query logs of the *Yandex* (2007) and *Excite* (2001) search engines. The *Yandex* is the

main Russian search engine and the *Excite* was a popular U.S. search engine in the early 2000s.

We study topic categories of queries related to Japan and corresponding categories of users submitting these queries. In this study, we compare two collective subjects: population of Russian and population of U.S. searchers. First of all, we will try to detect differences of search images of Japan among these populations. Another question is co-relation between searching different Japan-referring thematic classes.

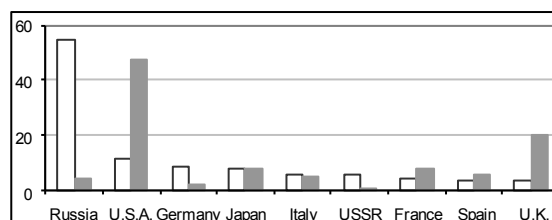


Fig. 1. Rates (%) of countries among 9 states referring to in Yandex-07 (white) and Excite-01 (gray) logs

We process the Japan-referring queries. While any country may be chosen as a perceived object, the reasonable questions are: is a number of queries referring to a country sufficient for statistically significant conclusions and how frequent are these queries among the queries related to other countries in the compared logs? We compare the rates of queries directly referring to 9 countries (Fig. 1). Fractions of Japan-referring queries are big enough and approximately equal in both logs.

II. RESEARCH QUESTIONS

The subject of the study is a *user* i.e. a set of all queries submitted by him rather than a *query* or a (task or temporal) *session*. In the study, we investigate fractions of Russian and U.S. users submitting Japan-referring queries. The same user may submit queries belonging to different Japan-referring classes. The research questions are:

- How frequently do Russian (both *Yandex* logs) and U.S. searchers submit queries of Japan-referring classes?
- How do Japan-referring classes co-occur in a set of queries submitted by the same Russian or U.S. user?
- A comparison of fractions and classes' co-occurrence of Russian and U.S. Japan-referring searchers (*Yandex-07* vs. *Excite-01*, a cross-time cross-nation comparison)

III. DATASETS

In this study we use 24-hour query logs of the Russian-language *Yandex* search engine (March 2007, 890,897 users) and 24-hour log sample of the U.S. *Excite* (May 4, 2001, 305,360 users). The datasets are different:

(1) *in time* (2007 vs. 2001; a “search image” varies over time). Since observation periods of the both datasets are equal to 24 hours we can ignore intraday variations. However, we cannot ignore week and season variations (e.g. in tourist queries) and we cannot ignore a long-term trend, which is particularly important and reflects changes of interests and change of the available Web services as alternative ways to get Japan-referring information. The datasets which we use are spaced far apart in time and search images may be shifted considerably over the years. As a result, can the comparison be valid? This is a crucial question for any, even time-synchronized, comparative study. Ideal comparative study of parallel social processes should be study of time series rather two time slices, even made in the same time.

(2) *in audience* (mainly Russian vs. mainly U.S.) and in population structure. Of course, not only Russians use *Yandex* and not only U.S. searchers used *Excite* in 2001. However, we can suppose that a majority of the *Excite-2001* queries were submitted by U.S. searchers because (1) about 90% of queries are in English and (2) only 11% of queries are submitted during “American day” (0am – 6am, Pacific time zone) when non-American users are active.

(3) *in language* (mainly Russian in the *Yandex* log vs. mainly English in the *Excite* log). The queries submitted to the *Excite* are queries in English (~90%), German and Spanish. The queries submitted to the *Yandex* contain words in Russian, two other Slavonic languages (Ukrainian and Belorussian) and in English. The *Yandex* users commonly use English spelling of Japanese brand names.

IV. JAPAN-REFERRING VOCABULARY CREATION

To detect and categorize Japan-referring words, queries and users we use two crucially different kinds of categories:

(1) *basic categories* corresponding to *both* aspects (a general reference to Japan and a certain thematic denotation, e.g., culture) and

(2) two *subsidiary categories-filters* used to detect those Japan-referring queries, which cannot be classified by perfect theme. These subsidiary categories are *general* (corresponding queries contain *japan** stem, e.g. <Japan>, <Japanese culture>) and *geography* (Japanese geographic and administrative names). Queries attributed to subsidiary categories should be categorized into basic categories in the next steps.

To detect and categorize Japan-referring queries we use the automatic procedure based on the Japan-referring vocabulary (hereafter only words from this vocabulary are referred to as “vocabulary words”). It contains both Russian and English words related to Japan. About 300 word-combinations, words and stems were selected (both Russian and English spelling for each word; and some words in each language were presented in different writings, e.g. *mitsubishi* and *mičubisi*). Table 1 exemplifies initial categories of Japan-referring words used in the preliminary analysis. Some words were attributed to multiple categories during the preliminary analysis.

TABLE I. EXAMPLES OF JAPAN-REFERRING WORDS

Category, Number of Words and Word-Combinations in	Examples
Subsidiary Categories	
<i>General</i> 17	Japan, Japanese, Nippon, Nihon
<i>Geography</i> 107	Chugoku, Tokyo, Kyoto
Basic Categories	
<i>Religion & ethic</i> 50	satori, shinto, tsukuyomi, zen, todaiji
<i>Traditional art, theater</i> 55	hokusai, netsuke, origami, utamaro
<i>History & interstate relations</i> 85	edo, hojo, meiji, samurai, taisho, tokugawa, yamato
<i>Traditional lifestyle</i> 45	kimono, ryokan, tatami, yakuza
<i>Literature</i> 37	haiku, kanji, mukai, renga, kobo abe, miyamoto musashi
<i>Masscult, movies</i> 16	anime, manga, pokemon
<i>Martial art</i> 24	aikido, budo, judo, karate, kendo, kyudo, sumo
<i>Traditional food</i> 26	sake, sashimi, sushi, tsukemono
<i>Cars</i> 30	mazda, tyota
<i>Consumer Goods</i> 59	Marubeni, canon

A serious problem in the query processing is a lot of typos and a confusing spelling. For example, while a Russian spelling of *Mitsubishi* is *Mitsubisi*, 433 Russian searches type Russian *Mitsubisi*, 51 searchers use Russian *Mitsubishi* (and 909 Russian searchers type *Mitsubishi* in English). To avoid a confusable spelling problem we use all probable variants of spelling.

Multi-categorization of vocabulary words. The original categorization allows a multi-valued word attribution, e.g. *kotatsu* belongs to both *religion* and *lifestyle* categories. However, since one of our goals is a study of cross-category dependency among all queries submitted by a user, this manifold word attribution is undesirable since it leads to artifactual detection of cross-category dependencies. Some

words have only one sense in any occurrence but senses of different occurrences are different. For example, *Hiroshima*, *Nagasaki* may occur either as historic or geographic terms, while *Pearl Harbor* is also a 2001 movie and a lot of *Excite*-2001 queries refer to the movie and a big fraction of historic queries is provoked by the movie.

V. PRELIMINARY ANALYSIS

The aims of the preliminary analysis are (1) a rough detection of categories among users’ queries, (2) disambiguation of variants of words use, and (3) detection of necessity and possibility to combine different categories into non-overlapping classes. There are 2 reasons to combine different categories: (a) irremovable co-occurrence of different categories for some words and (b) too small rates of several categories among queries.

Each query are attributed to all categories of the vocabulary words contained in the query and a user is attributed to all categories of the vocabulary words contained in all queries submitted by him. Users submitting queries containing multi-attributed vocabulary words are attributed to all categories of these words. 29,208 (3.28%) of 890,897 *Yandex* users and 4,553 (1.49%) of 305,360 *Excite* users submitting queries containing the Japan-referring words. Fig. 2 shows the distributions of the *Yandex* and *Excite* users among Japan-referring categories.

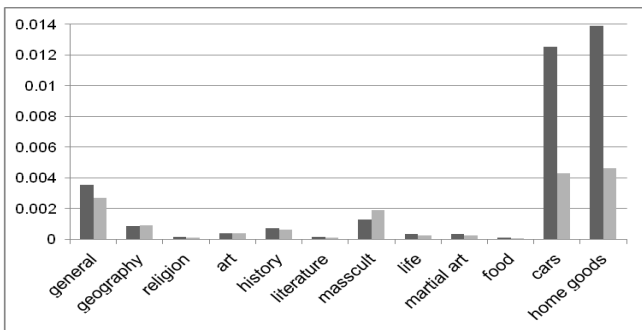


Fig. 2. Rates (%) of categories of Japan-referring users in the Yandex (black) and Excite (gray) logs

Table 2 shows number and fractions of users submitting queries containing words of corresponding categories. To compare these fractions in the *Yandex* and *Excite* logs we use z-test in form of:

$$z = \frac{|\hat{P}_{Yandex} - \hat{P}_{Excite}|}{\sqrt{\hat{p}(1 - \hat{p})(1/n_{Yandex} + 1/n_{Excite})}} \quad (1)$$

where p_{Yandex} and p_{Excite} are sample rates for the category in each log and p is a sample rate in a combined population. Fractions of consuming categories (*cars*, home *electronics*, *other consumer goods*) and *masscult* category are enormously different. Fractions of two categories (*history* and *literature*) are different for $z_{0.95} = 1.96$ but we cannot discard the null hypothesis at $z_{0.99} = 2.58$. Fractions of *geography*, *religionðics*, *arts* and *traditions* categories are equal even at 0.95.

TABLE II. COMPARISON OF FRACTIONS OF USERS ATTRIBUTED TO CERTAIN JAPAN-RELATED CATEGORIES AMONG 29,208 *YANDEX* AND 4,553 *EXCITE* USERS SUBMITTING JAPAN-RELATED QUERIES

Category	Yandex users		Excite users		z test
	Number	Fraction (%)	Number	Fraction (%)	
general	3,158	0.355	820	0.269	7.12
geography	812	0.091	272	0.089	0.33
religionðics	125	0.014	35	0.015	1.06
arts	247	0.028	71	0.023	1.31
traditions	120	0.013	43	0.014	0.25
history&interstat	643	0.072	185	0.061	2.10
e literature	159	0.018	37	0.012	2.14
masscult,movies	1,152	0.129	585	0.192	7.80
life	328	0.037	74	0.024	3.27
martial art	312	0.035	76	0.025	2.69
meal	112	0.013	20	0.066	2.73
cars	11,160	1.253	1,308	0.428	38.71
electronics	10,485	1.177	1,026	0.336	41.08
other goods	1,920	0.216	383	0.125	9.80

VI. RE-CATEGORIZATION: COMPOUND CLASSES

We take into account the preliminary analysis results regarding (1) a size of users categories (size of some categories is small for statistical inferences) and (2) a ambiguous categorization:

(a) closely related categories are combined into compound classes. The resulted 7 classes (5 basic classes and subsidiary *general* and *geography*) are shown in Table 3. We do not change attributes of the vocabulary words assigned in terms of 12 initial categories. Only processing is changed: if a word belongs to any category it accounted as belonging to corresponding class.

(b) vocabulary words belonging to different *new* classes are re-attributed to avoid a multiple categorization in terms of classes. (The only exception is *Pearl Harbor* which is frequently used both as *masscult* (the movie) and as *history*. Queries containing *Pearl Harbor* are attributed manually either to *masscult* or to *history*). Since some words and queries initially attributed to *geography* and *history* were re-attributed, data in Table 4 differs from data in Table 3. Now, if a user attributed to several classes, queries of this user really contain different words belonging to these classes.

TABLE III. NON-OVERLAPPED CLASSES OF WORDS

Class	Categories included into Class
<i>general</i>	general
<i>geography</i>	geography
<i>culture</i>	religionðics, arts, traditions, literature, life, food
<i>history</i>	history & interstate_relations
<i>martial ort</i>	martial art
<i>masscult</i>	masscult, movies
<i>goods</i>	cars, home electronics, other consumer goods

Fig. 3 and Table 4 show rates of users submitting queries containing reclassified Japan-referring words. Now all fractions are different at $z_{0.95}$ for Russian and U.S. searchers (cf. Table 2). Fractions of all classes among *all* Russian searchers are bigger than corresponding fractions among *all*

U.S. searchers. At the same time, fractions of all non-consuming classes among *Japan-referring* searchers are significantly smaller than corresponding fractions among U.S. *Japan-referring* searchers. The Russian *Japan-referring* search is mainly consuming.



Fig. 3. Rates (%) of classes of the Japan-referring users among all Yandex (white) and Excite (gray) users

TABLE IV. COMPARISON OF FRACTIONS OF USERS ATTRIBUTED TO CERTAIN JAPAN-RELATED CLASSES AMONG 29,208 YANDEX AND 4,553 EXCITE USERS SUBMITTING JAPAN-RELATED QUERIES

Category	Yandex users		Excite users		z test
	Number	Fraction	Number	Fraction	
<i>general</i>	3,158	0.355	820	0.269	7.12
<i>geography</i>	687	0.077	158	0.052	4.55
<i>culture</i>	890	0.100	204	0.067	5.22
<i>masscult</i>	1,152	0.129	585	0.192	7.80
<i>history</i>	606	0.068	174	0.057	2.06
<i>martial art</i>	312	0.035	76	0.025	2.68
<i>goods</i>	23,029	2.585	2,663	0.872	56.35

Co-occurrence of classes. If a user submits queries of several classes he is attributed to all these classes (“multi-class user”). A few users were attributed to two classes as a maximum. Table 5 shows the contingency table of users automatically attributed *only to basic classes* (as a result, diagonal values in Table 5 are less than values in Table 4 since users attributed to subsidiary classes are frequently attributed to other classes too); users attributed (also) to *general* and *geography* subsidiary classes are not included and will be re-attributed in the next chapter.

TABLE V. CONTINGENCY TABLE FOR 25,332 YANDEX AND 3,591 EXCITE “NON-SUBSIDIARY” USERS

Yandex	culture	history	martArt	masscult	goods
history		576	2	10	37
martial arts			296	1	11
masscult				1,116	15
goods					22,705
Excite	culture	history	martArt	masscult	goods
culture	183	1	0	5	2
history		158	1	5	10
martial arts			74	0	0
masscult				564	2
goods					2,638

VII. RE-ATTRIBUTION OF SUBSIDIARY CLASSES

Now we should classify users who were automatically recognized as belonging to two subsidiary classes (*general* and *geography*). The rates of these users are big enough (Table 4) and they should be automatically or manually categorized into basic classes.

The idea is that non-vocabulary words which co-occur with a vocabulary word may be related to the basic class assigned to this word. Two types of the co-occurrence were considered: (1) a *narrow* query-based co-occurrence in the same query and (2) a *wide* user-based co-occurrence in a whole set of queries submitted by the same user (e.g., if a user attributed just to one class *martial art* submits two queries *<tortie cat>* and *<jujutsu>*, then words *tortie* and *cat* co-occur with *jujutsu* and are considered as possible associated words of the *martial art* class). The first step of the automatic classification is mining of *non-vocabulary* words associated with any class. To mine them we use only those items (Japan-referring queries in the case of the narrow co-occurrence or Japan-referring users in the case of the wide co-occurrence), which are attributed just to one class. Next, if extracted co-occurred words more frequently occur in Japan-referring queries (a narrow co-occurrence) or in any query of Japan-referring users (a wide co-occurrence) attributed to the class these co-occurred words are considered as associated non-vocabulary words of this class. Let associated words be extracted for each basic class. Then queries attributed to subsidiary classes may be re-attributed to basic classes by occurrence of associated words.

To extract the non-Japan-referring terms, which represent the classes of queries we use following class-based metrics:

— $tf(term\ T\ | \ item_of_class_Cl)$ — “class frequencies” of the non-vocabulary *term T* in *item_of_class_Cl*, i.e. the ratio of the number of *term T* occurrences in *item_of_class_Cl* to the total number of all words occurrences (a total length) in all unique queries belonging to *item_of_class_Cl*. The *item_of_class_Cl* is either any query containing vocabulary words belonging to *class_Cl* (the narrow co-occurrence) or any query of a user submitting at least one query containing vocabulary words belonging to *class_Cl* (the wide co-occurrence).

— $cf(term\ T)$ — “collection frequency” of the non-vocabulary *term T* in all unique queries of the query collection, i.e. the ratio of the number of *term T* occurrences in all unique queries of the query collection to the number of all words occurrences in all unique queries of the collection.

— $contrast(term\ T\ | \ item_of_class_Cl) = tf(term\ T\ | \ item_of_class_Cl) / cf(term\ T)$. If this ratio is significantly bigger than 1, then *term* represents *class_Cl*.

We count $contrast(term\ T\ | \ class\ Cl)$ to detect non-vocabulary words closely connected to the vocabulary classes: terms which occur either (1) in the queries belonging to the class more often than to the other queries (narrow co-occurrence) or (2) in all queries submitted by users attributed

to the class more often than in the queries of all other users (wide co-occurrence).

Fig. 4 shows results of re-classification of 3195 (of 3845) Russian and 703 (of 955) U.S. searchers primarily recognized as subsidiary classes (other “subsidiary users” submitted too general queries such as <Japan> were not re-classified). Fractions of *sex* and *cars* classes are enormously different and look like mirror images: At first sight it may be interpreted as a result of the *from-e-sex-to-e-commerce* tendency (Spink et al., 2002a). However we do not discover this tendency in *Japan-referring* queries comparing the *Excite* logs (2001 vs. 1999) or *Yandex* logs (2007 vs. 2005). Too low fraction of Japan-referring sex searchers among *Yandex* users may be partly explained by the fact that *Yandex* covers only the Russian Web domain.

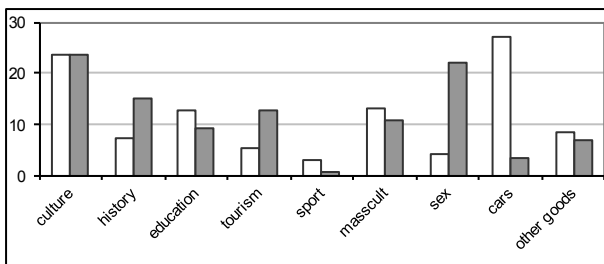


Fig. 4. Rates (%) of basic classes among Yandex and Excite users initially attributed to subsidiary classes

Even forth query in the *general* class is <Japanese autos>. Topics of *history* queries are really different between Russian and U.S. queries. While the latter are focused on World War II, the former practically ignore the WW II period but search for such topics as *Japanese ethnos forming* or *constitution of 1899*.

Let’s present “subsidiary” Japan-referring users in terms of basic classes to add these re-classified users to other Japan-referring users automatically attributed to the basic classes in the previous chapter. Namely, we combine *culture*, *education* and *tourism* classes into basic *culture* class, *sex* is added to *masscult*, *cars* and *restaurants* are added to *goods*. Since any user is attributed to two classes as a maximum, to group classes we use inclusion-exclusion rule for two sets:

$$n(\text{Group}) = \sum_{\text{class} \in \text{Group}} n(\text{class}) - \sum_{\text{class1}, \text{class2} \in \text{Group}} n(\text{class1} \cap \text{class2})$$

$$n(\text{Group1} \cap \text{Group2}) = \sum_{\text{class1} \in \text{Group1}, \text{class2} \in \text{Group2}} n(\text{class1} \cap \text{class2}) \quad (2)$$

Classes co-occurrence. Table 6 presents a contingency table of manually re-classified users initially attributed to *general* and *geography* subsidiary classes. Since not all queries containing words of subsidiary classes (*general* and *geography*) may be recognized in terms of 5 basic classes some of “subsidiary users” were not re-attributed.

TABLE VI. CONTINGENCY TABLE FOR 3,195 YANDEX AND 703 EXCITE MANUALLY RE-ATTRIBUTED SUBSIDIARY USERS

Yandex	culture	history	martArt	masscult	goods
culture	1,305	104	15	30	19
history		234	3	4	6
martial art			103	0	3
masscult				543	14
goods					1,198
Excite	culture	history	martArt	masscult	goods
culture	311	11	1	9	3
history		108	0	1	0
martial art			5	0	1
masscult				225	0
goods					79

VIII. CONSUMERISM vs. “MASSCULTURISM”

Now we can add re-classified “subsidiary” searchers to searchers classified by the five basic classes (*culture*, *history*, *martial art*, *masscult* and *goods*). Fig. 5 shows rates of 5 basic classes among users classified as *Japan-referring* rather among all users (all rates among *all users* of the search engine are bigger for Russians).

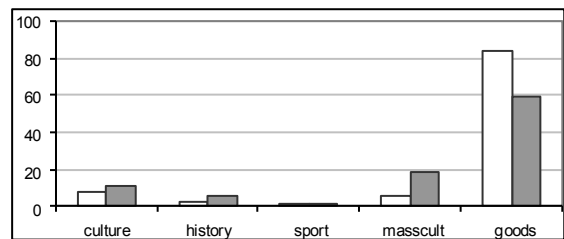


Fig. 5. Fraction (%) of users belonging to basic classes among Yandex and Excite Japan-referring users

TABLE VII. CONTINGENCY TABLE FOR 28527 YANDEX AND 4294 EXCITE JAPAN-REFERRING USERS

Yandex	culture	history	mArts	massc	goods
culture	2,100	121	26	43	58
history		810	5	14	43
martial arts			399	1	14
masscult				1659	29
goods					23903
Excite	culture	history	mArts	massc	goods
culture	494	12	1	14	5
history		266	1	6	10
martial arts			79	0	1
masscult				789	2

The combined categorization of manually and automatically classified users shows that Russian searchers demonstrate much more consumer interests. This is in accordance with the difference between U.S. “teenagers” and Russian “steadies” which is enormous among re-classified users. While we do not know age of the *Excite* users and only

partly know distribution of the *Yandex* population by age ([11]), we suppose the revealed difference of Japan-referring searches is not explained by the age difference between Russian and U.S. searchers.

IX. CLASSES CO-RELATION

Table 7 presents a final contingency table for all users attributed to basic Japan-referring classes, i.e. Table 7 is a sum of contingency tables of users automatically attributed to basic classes (Table 5) and re-attributed users (Table 6). How do basic Japan-referring classes co-occur in a set of queries submitted by the same user? To detect closely interrelated classes we estimate the probability of a random co-occurrence of classes among independent classes of Japan-referring users. Our goal is to detect such cases of intersections of classes that infract the assumption about independency of classes.

Let n_i be the number of users attributed to the class i (diagonal elements in Table 7), $obs(i,j)$ be the number of users attributed to both classes i and j (non-diagonal elements in Table 7), and N be the number of all considered users. To measure the strength of the interrelation between two classes we use a probably $p(k \geq obs(i, j))$ that a number of random co-occurrences k of independent classes i and j (containing n_i and n_j users) is not less than the observed intersection $obs(i, j)$. This measure shows to what extent the observed interrelation is incompatible with the assumption of independence of the classes. The smaller $p(k \geq obs(i, j))$, the stronger the interrelation is.

$$p(k \geq obs(i, j), n_i, n_j, N) = \sum_{k=obs(i, j)}^{k=\min(n_i, n_j)} p(k, n_i, n_j, N) \quad (3)$$

where $p(k, n_i, n_j, N)$ is a hypergeometric probability of k co-occurrences of n_i marks of the type i and n_j marks of the type j which are independently used to mark N "cells"

$$p(k, n_i, n_j, N) = \frac{\binom{n_i}{k} \binom{N-n_i}{n_j-k}}{\binom{N}{n_j}} = \frac{n_i! n_j! (N-n_i)! (N-n_j)!}{k! N! (n_i-k)! (n_j-k)! (N+k-n_i-n_j)!} \quad (4)$$

We consider probabilities of class co-occurrence among all users ($N_{Yandex}=890,897$ users, $N_{Excite}=305,360$ users). This approach to class co-relation is absolutely correct but is not very expressive. Indeed, we can a priori suppose the co-relation of Japan-referring classes. As a result, we can expect that probabilities $p(k \geq obs(i, j))$ that a number of random co-occurrences k of independent classes i and j is not less than the observed intersection are small. Rather, when we consider probabilities of co-occurrence among all users, the non-small probabilities are surprising and should be of special interest as "symptoms of independence" of classes.

We also consider probabilities of class co-occurrence only among the users attributed to Japan-referring basic classes ($N_{Yandex} = 28,527$ users, $N_{Excite} = 4,294$ users). While this opposite "over-strong" approach is surplus (in particular, it elaborates the same ordering of probabilities), it visualizes

differences between strong interclass relations. It is very expressive when we want emphasize the closest connections between classes, i.e. to differ strong co-relations (small probabilities of observed co-occurrence) from over-strong ("the smallest" probabilities). Tables 8, 9 shows estimations of probabilities $p(k \geq obs(i, j))$ that a number of random co-occurrences of independent classes is not less than the observed intersection for both considered sets of users.

1. "Independence criterion" (probabilities of Japan-referring class co-occurrence among all searchers, Table 8). While *goods* and *masscult* classes, at first sight, should be more co-related than, for example, *goods* and *martial art*, these biggest classes are practically independent in both audiences (more than 0.99 probability of a random co-occurrence). In general, *goods* class is the most independent in both audiences. The only exception presents the Russian audience for which *goods* and *history* are strong co-related classes (in contrast with independence of these classes in the U.S. audience). Co-relations of the *martial art* significantly differ among audiences: this class is more co-related with *goods*, *masscult* and partly *culture* classes in the Russian audience (0.189 vs. 0.506, 0.524 vs. 1 and 0 vs. 0.12

TABLE VIII. PROBABILITIES OF THE RANDOM CLASS CO-OCCURRENCE AMONG ALL USERS (INDEP. CRITERION)

Dataset	890,897 <i>Yandex</i> users, 305,360 <i>Excite</i> users			
<i>Yandex</i>	history	martial arts	masscult	goods
culture	0	0	0	0.42974
history		0.00004	0	0.00003
martial arts			0.52473	0.18939
masscult				0.99503
<i>Excite</i>	history	martial arts	masscult	goods
Culture	0	0.12008	0	0.64105
History		0.06654	0.00008	0.90730
martial arts			1	0.50646
masscult				0.99305

2. "Over-strong co-relation" criterion (probabilities of Japan-referring class co-occurrence among Japan-referring searchers, Table 9) reveals the big difference between a strong co-relation of the *culture* and *history* classes in U.S. audience and the *strongest* co-relation of these classes in the Russian audience (0.999 vs. 0).

TABLE IX. PROBABILITIES OF THE RANDOM CLASS CO-OCCURRENCE AMONG JAPAN-REFERRING USERS

Dataset	28,527 <i>Yandex</i> users, 4,294 <i>Excite</i> users			
<i>Yandex</i>	history	martial arts	masscult	goods
culture	0	0.76908	1	1
history		0.98917	1	1
martial art			1	1
masscult				1
<i>Excite</i>	history	martial arts	masscult	goods
culture	0.99999	0.99994	1	1
history		0.99390	1	1
martial art			1	1
masscult				1

On all occasions, interdependency between classes is stronger in the *Yandex* audience and this is not an artifact.

X. CONCLUSION

We have investigated (1) differences between Russian and U.S. search images of Japan and (2) interdependency between searching for different Japan-referring classes: how frequently searchers of one Japan-referring topic also search for other topics.

1. Fractions of all classes *among all* Russian searchers are bigger than the fractions *among all* U.S. searchers. At the same time, fractions of the non-consuming classes *among Japan-referring* Russian searchers are significantly smaller than fractions *among Japan-referring* U.S. searchers. The Russian Japan-referring searchers are mainly consuming, whilst the U.S. Japan-referring searchers are much more masscult-oriented. A fraction of culture-oriented searchers is small in both audiences.

2. The Japan-referring *goods* class primary relate to goods rather than to Japan, and users submitting Japan goods queries do not frequently submit other Japan-referring queries. On the contrary, the *masscult* class is compatible with non-consuming classes and surprisingly is not compatible with *goods* in both audiences and *sport* in the Russian audience.

3. The Russian searchers submitting queries referring to Japan culture relatively frequently submit other Japan-referring queries, especially queries related to the history of Japan. Furthermore, all Japan-referring classes are more co-related in the Russian audience.

REFERENCES

- [1] S. Beitzel, E. Jensen, A. Chowdhury, D. Grossman, and O. Frieder, "Hourly analysis of a very large topically categorized Web query log", in proceedings of 27th ACM SIGIR conf. on research and development in information retrieval, 2004, ACM Press, pp. 321-328.
- [2] S. Beitzel, E. Jensen, A. Chowdhury, O. Frieder, and D. Grossman, "Temporal analysis of a very large topically categorized Web query log". J. of the Association for Information Science and Technology, vol. 58, No. 2, 2007, pp. 166-178.
- [3] D. Blei and J.D. Lafferty, "Dynamic topic models", in Proceedings of 23rd Int. Conference on Machine Learning ICML (Pittsburg, USA, June 2006), ACM Press, pp. 113-120.
- [4] B.J. Jansen, A. Spink, and T. Saracevic, "Real life, real users, and real needs: a study and analysis of user queries on the Web". Information Processing & Management, vol. 36, no. 2, 2000, pp. 207-227
- [5] D. Lewandowski, "Query types and search topics of German Web search engine users", Information Services&Use, vol. 26, 2006, pp. 261-69
- [6] M. Richardson., "Learning about the World through Long-Term Query Logs". ACM Trans. on the Web, vol. 2, no. 4, 2008, pp 21-27
- [7] A Spink., S. Ozmutlu., H.C. Ozmutlu., and B. Jansen, "U.S. versus European Web searching trends", ACM SIGIR Forum, vol. 36, no. 2, 2002, pp. 32-38
- [8] X. Wang and A. McCallum, "Topics over time: a non-Markov continuous time model of topical trends", in Proceedings of KDD '06 (Philadelphia, USA, August 2006), ACM Press, pp. 138-145
- [9] I. Weber, V. Garimella, and E. Borra, "Mining Web Query Logs to Analyze Political Issues", in proceedings of the WebSci 2012 conference, June 22-24, 2012, Evanston, Illinois, USA, ACM Press, 2012, pp. 330-334.
- [10] [US-to-Japan-Polls] (2014 and earlier) The U.S. Polls on opinions toward Japan. <http://www.mofa.go.jp/region/n-america/us/survey/index.html>
- [11] [RU-Net] (2014 and earlier) Project "The internet in Russia/Russia on the internet".

Internet-based Troubleshooting and Monitoring System of Industrial Robots

Md Mozasser Rahman¹

Sulleha Bt Parnin²

Department of Mechatronics Engineering
International Islamic University Malaysia (IIUM)
Kuala Lumpur, Malaysia

¹mozasser@iium.edu.my, ²sullehaparnin@gmail.com

Abstract—This paper presents a prototype of an industrial robot monitoring system that is remotely control via internet. The use of industrial robot is increasing in small to medium enterprise (SME's) as low-cost robots are available in market. Productivity can be increase by incorporating robots. The limitations of maintenance and troubleshooting personnel to monitor the robot will always exist because they are not always being in the robot operation area. Thus, a prototype of a monitoring system was developed and tested in the Robotics laboratory of International Islamic University Malaysia, using Denso robots. The remote monitoring system can monitor the robot condition by a real time video streaming using internet. Arduino microcontroller was used to send the signal to the robot from a computer. This open source software fit the requirement of this project which enable the development to be completed with a low cost budget. The user interface program was created using the Guided User Interface which enables the programmer to monitor the robot from a remote area by using TeamViewer application. A USB camera was also used to see the real time video of the robot. All the subsystem was successfully integrated by Define, Data Gathering and Analysis, Develop, Testing and Improve to accomplish the project goals and at the same time becoming a starting point for further development.

Keywords—Industrial Robots; troubleshooting; monitoring; Internet-based; SME, Lowcost.

I. INTRODUCTION

The productivity of Small and Medium-sized Enterprise (SME) can be increased by using robots. Almost all of the big industrial robot makers have or are working on lightweight and human-friendly arms, but none are offering them at low cost, or with user-friendly training, or the plug and play and safety features. There are other robot startups in the SME marketplace – and on the horizon – but none are as far along in their development and low cost as **Rethink Robotics** and **Universal Robots** [1]. Some of these SME need only one or two robots for the automation process. It is not feasible for them to recruit a programmer or another skilled technician for troubleshooting those robots. There are after sales service provided by the suppliers but down time will be high. Thus, a prototype of a monitoring system was developed and tested in the IIUM Robotics laboratory by using the Denso robots. The remote monitoring system can monitor the robot condition by a real time video streaming using internet with a low cost budget. Arduino microcontroller was used to send the signal to the robot from a computer.

Marin et. al. developed user interface program for remote programming the robot via internet. Programming language is JAVA [2]. Rosado-Muñoz also proposed a web-based remote laboratory for the programming of the industrial robot. A webcam use to provide a video of the robot movement. Both are for educational purposes [3]. 3D modeling of a robot for physical assembly using internet application server for image processing has been developed by Wang et al. [4]. Their experimental result shows that the system is feasible to meet

industrial assembly by producing a shorter image processing time.

Interest in the design of Internet-based telelaboratories is increasing enormously, and this technique is still very new. Most of the systems are cost oriented and need a special training [5-6]. Only a low cost and user-friendly system provide the realistic solution for SME's. Therefore an user interface program was created using the GUI which enable the service provider to monitor the robot from a remote area by using TeamViewer application.

II. SYSTEM OVERVIEW

A six-axis articulated robot as shown in Fig. 1 was used as a target robot. The proposed system consists of four subsystems which are:-

- Subsystem 1: Robot controller and server PC interfacing.
- Subsystem 2: GUI for communication.
- Subsystem 3: Camera monitoring.
- Subsystem 4: Interfacing client and server computers.



Fig. 1. Denso robot, used for the prototype.

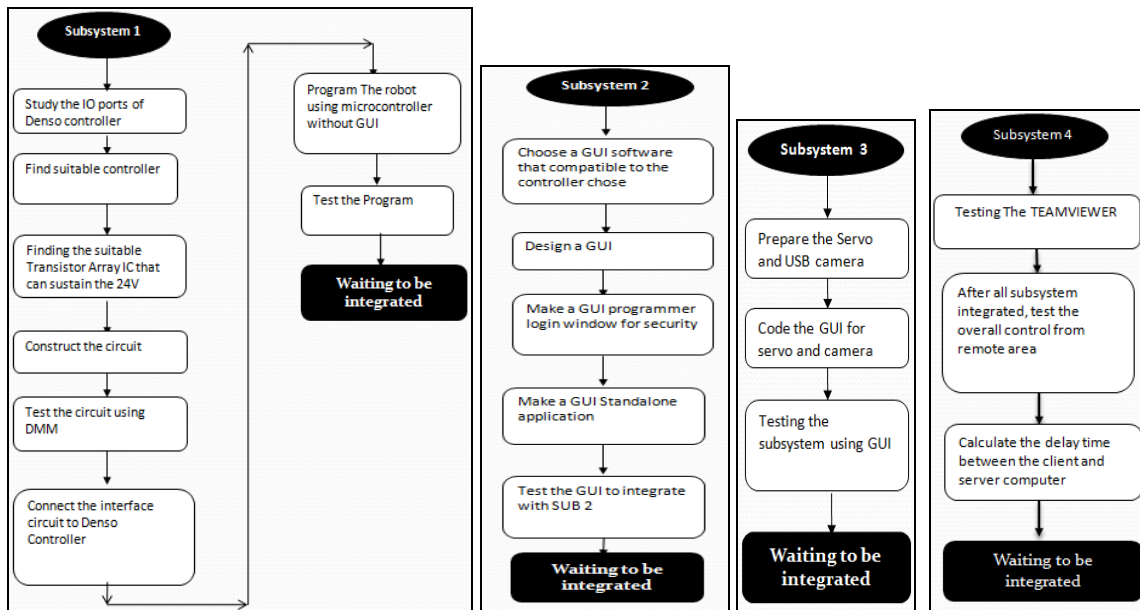
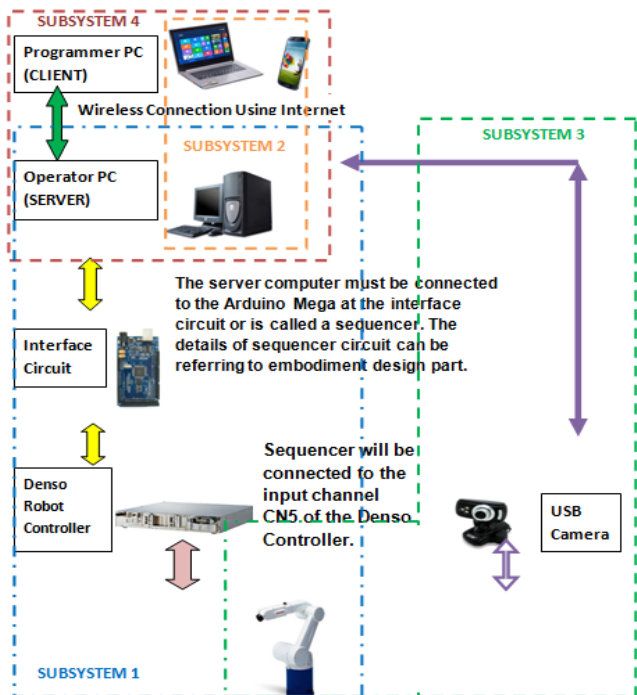


Fig. 2. Subsystem development flowchart

All the systems were developed simultaneously with priority given to subsystem 1 and 2. The subsystem development flowchart is presented in Fig. 2. Robot controller and server PC interfacing dominate the system.

Fig. 3 shows the position of the subsystem in the main system. All subsystems are mutually dependent. The arrow in the figure indicates the relationship of signal transferred.

After indentifying the function of the Denso controller i.e to the IO channel of the controller. This input channel is the one



that is connected to the digital pin of the Arduino. There are a total of 24 CN5 pins that successfully been connected to the Arduino. By using the GUI, the signal status (high or low) of these pins can be controlled.

III. WORKING PRINCIPLE

In every system, working principle describe the rule that enable the system to give a desired output. In this case, the input is the command from the programmer that give output in terms of robot motions, sensor data, robot image and etc. This input and output relation is shown by the system of work principle in Fig. 4. It applies that, programmer to sit at the client side and operator/supervisor will be at the server side. This system is powered by Matlab 2013-a, Arduino 1.0.6, and Teamviewer 10.

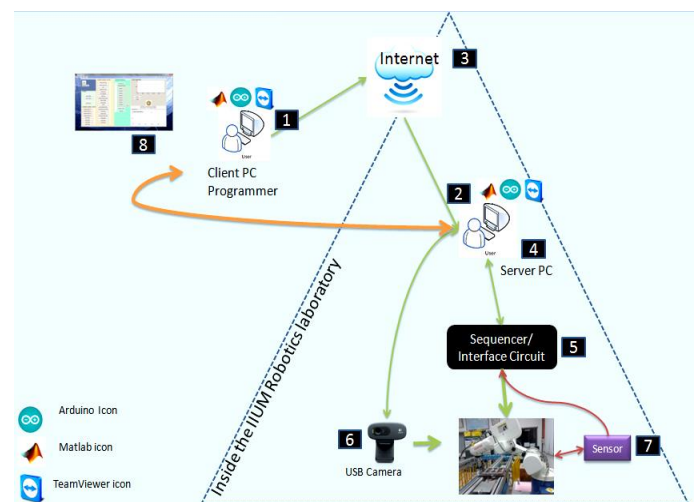


Fig. 4. Working principle of the system.

According to Fig. 4, the triangle in dashed refers to the testing area which is the robotics laboratory. The numbers in the figure can be described by the following:-

1. Programmer sign-in to TeamViewer account. Programmer will then need to connect to the server PC to enable the access to the robot. Programmer will login to the system and change the mode to external mode of operation. It is required that the Client computer (programmer PC) to have Arduino IDE and Matlab software.
2. The server computer must be turned on all the time. TeamViewer applications must be open to enable the access from the client PC. Server computer need to have Arduino IDE and Matlab software as well. The USB of the interface circuit and camera must be connected to the USB ports of the server PC. This is to ensure the direct signal transfer from the GUI to the circuit. An Arduino code to make it as server is uploaded during the subsystem 1 development. Therefore, Arduino code was written inside the Matlab code itself as the Arduino is now become a server that can only received a data from Matlab-GUI through serial port.
3. TeamViewer will take some time to verify the IP address of the server computer. Once the access is successful, programmer can view the server PC as if they were sitting in front of it.
4. Operator or supervisor that is in the robot operational area still can get control over the robot since the mode is INTERNAL. Once the programmer runs the program sequential in the monitoring window, operator/supervisor will not be able to run the robot by teaching pendant. It indicates the mode change from INTERNAL to EXTERNAL. Meanwhile, at the programmer side, they can start to open the Login window of the system and start to take over the robot operation using the user interface monitoring window.
5. Any command from the programmer will be received by interface or sequencer circuit.
6. The motion of the USB camera can be controlled by adjusting the servo position either to the left or to the right. Those two positions will view two different robots. The mechanical part for the camera was design to be able to monitor two robots. A high definition USB camera was used to ensure the quality of the robot video and images. This camera monitoring system covers subsystem 2 and 3.

7. The additional feature of the system is to put a sensor that can detect the distance of the object (for the case that robot was design for a pick and place operation with a conveyer system). In this project, a SHARP IR obstacle detection sensor was used with the capabilities to detect object at most 80cm.
8. Data such as the status of the IO pins, sensors value is accessible through Matlab-GUI with some delays in feedback. Delays can be depending from the speed of the internet connection of both server and programmer side and any other factor such as the computer CPU performances.

The principle of work in this project cannot be accomplish without knowing the interface circuit as it is the only medium to connect the computer to the Denso robot controller. Fig. 5 shows the connection done on the interface circuit. The main components of the circuit are the Arduino Mega and transistor array ULN 2803. According to Fig. 5, three transistor arrays were used to drive the signal from the GUI to the Denso robot controller. It enables the signal to be passing with a minimal input current. The advantage on using transistor array is it can simplify the circuit complexity. It composed of multiple transistors in one package and helps to improve mounting density and reduce the board population costs associated with the use of discrete transistors.

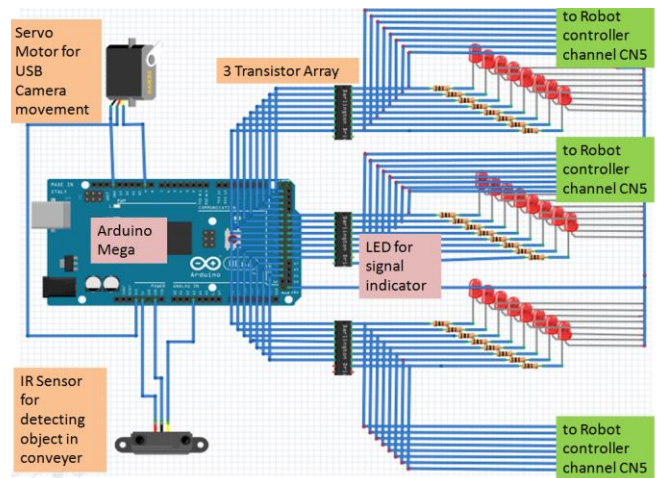


Fig. 5. Circuit connection for one robot attachment.

IV. RESULT

The following steps and its respective figure are the result of integrating the whole system.

Preparation at the server PC side before the system fully tested:

- i. Power on the robot. Robot power on at this time indicates the INTERNAL mode of operation.

- ii. Open the Arduino>examples>pde>Adioes. Upload the Adioes code to the Arduino board. This is to be done only once to make the board as the server. All the instructions for the board input and output pins was done in GUI-Matlab editor file.
- iii. Open the Matlab. Matlab should not be open directly from the main icon. It must be open through the folder of 'MatDuino'.

Meanwhile, preparation at the programmer side is for the programmer to sign in to the TeamViewer account and connect to the server PC. After the access successful, programmer can initialize the system by running the Matlab GUI name 'Login'. Upon running the login, provide the details need in login window such as the programmer ID and password as in Fig. 6. If login is successful, the main monitoring window will appear. If else, user cannot be allowed to enter the main monitoring window. It works like the security for the system from the stranger.

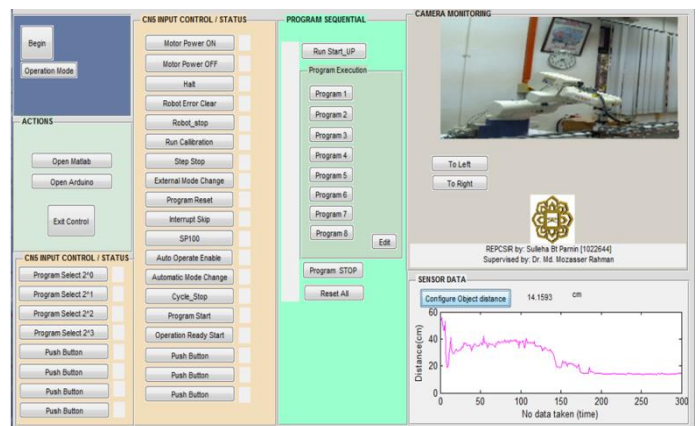


Fig. 7. GUI window for incomplete subsystem integration (only the sensor is tested)

Monitoring window with the name 'cc' at a moment after the login is successful.

Fig. 8. GUI after imp

on the robot can be monitored from the remote location.

To begin, click on the pushbutton name 'Begin' to allow control of the robot. The Mode of operation must be in the 'INTERNAL' mode. This is because the startup operation is still not executed. Note that when the 'Startup' button is pushed, operation mode will automatically change to 'EXTERNAL'. Fig. 7 shows the feature of the monitoring windows.

Monitoring window was tested and is improved in terms of programming of the speed of the camera and the status of process in the program sequential panel. Fig. 8 shows the 'EXTERNAL' mode of operation for the controlled robot while the camera was monitoring another robot operation.



Fig. 6. Screenshot of login page



V. DISCUSSION

Running in the External mode of operation will change the status of the CN5 start-up input pins of the Denso robot controller from low (white) to high (green). Green color on the status panel is also indicates the signal is active and being sent to the robot controller through the interface circuit. Startup operation caused the mode to be change to automatic, motors to turn on, and calibrated the robot position.

The startup step is shown in Fig. 9. Program can be run right after the startup process is finished which indicate by the 'EXTERNAL' in the operation mode. Simultaneously, the camera and sensor can be tested.

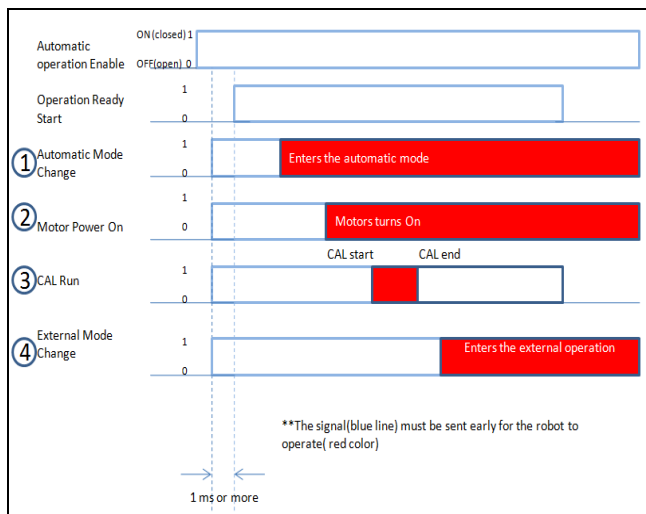


Fig. 9. Robot Startup steps (Signal Operation diagram)

The Digital pins involves are 24(Auto Operate Enable), 48(Automatic Mode Change),43(Operation Ready Start), 44(Motor Power On),46(Run CAL), and 52(External Mode Change). Firstly, digital pin 24 need to be set high . This is followed by pin 48, 44, 46, and 52 and 43. This signal is an input to the Denso controller. It is shown by the blue line of the diagram in Fig. 9.

According to Fig. 9, the output response of the startup is shown by the red colour. Even though the signal is high at the same time, the reponse need to wait for the successfulness of the previous start up steps to initiated. This means that the external mode will not be achieved unless the calibrationof the motor is executed and completed. Internal mode light will be off to indicates that the external mode is turned on. Since the robot is in the external mode, no other instruction from the teach pendant will be executed. Therefore, only the instruction from the Arduino will be allow to operate the robot. In other word, this system will work only when the robot is in EXTERNAL mode of operation.

It is found that, the duration of startup time for the robot is less than that needed to startup using a teach pendant. The motor is turned on in much faster as the system eliminate the human command delay of pressing teach pendant buttons. If the time taken for a student to program the robot startup operation is compared with the time taken for the system startup operation, the system will approximately faster at a rate of two times the teach pendant. Testing the overall system require about a maximum of 10 minutes just to connect the MatlabIO to the serial port of Arduino. The attempt for Matlab to connect to Arduino is always fail required a special line of code to be executed on the Matlab command window which will delete the currentport and force the release at the serial connection. The following code was used:

- i. To delete the current MATLAB serial connection on COM26 : `(instrfind({'Port'},{'COM26'}));`
- ii. To delete all MATLAB serial connections : `delete (instrfind('Type', 'serial'));` .

While the overall system is tested, the status of each pin of the input pin CN5 can be checked by the red color to indicated the signal is being sent as 'HIGH' or '1' to the particular pin. These features enable the programmer to detect the problem whether any of the pins were caused an error to the robot operation. Programmer also may change the mode of operation from the 'EXTERNAL' operation to the 'INTERNAL' operation by pressing the pushbutton name 'Reset All'. Then all the digital output pins will become low as there were no signals being allowed to pass the sequencer circuit.

Through the CN5 cable of the Denso robot controller, robot fault can be cleared automatically without using the teach pendant in the EXTERNAL mode of operation. The 'High' combination of signal of 'robot error clear (CN5-18) and 'Operation Ready Start (CN5-23) will cause the error of the robot to be cleared. CN5-23 must be set into high at least 1 milliseconds to allow the robot to prepared for the motion and to prevent errors.

VI. CONCLUSION

An Internet based Troubleshooting and monitoring system of industrial robot was successfully developing. It is achieved by design and implementations of the interface circuit. Result shows that signal send by Arduino can be received by the Denso Controller perfectly similar to the teach pendant. The server computer also can be accessible by the client computer by using an internet connection although the quality of the controlled windows is not good enough. Finally, the programmer that is in remote area is able to see the robot status by running some startup operation on the robot. This

system will surely can be implemented in a medium or small robotics industry. The controlling of the robot in this way actually mimic the control system in a large organization which to implement a SCADA or DCS.

VII. RECOMMENDATION

For future improvement of the project, additional features of the GUI system can be added such as the error elimination, motor gripper control, and getting the current angle of the robot actuator by using a feedback sensor. Also, it is found that, one Arduino Mega can control up to two Robots and one computer can detect for more than one Arduino COM port. This makes us able to control many robots just in one computer wirelessly. It can be said that the connection made between the server computers to the robot is similar to the Profibus application as we monitor the robot.

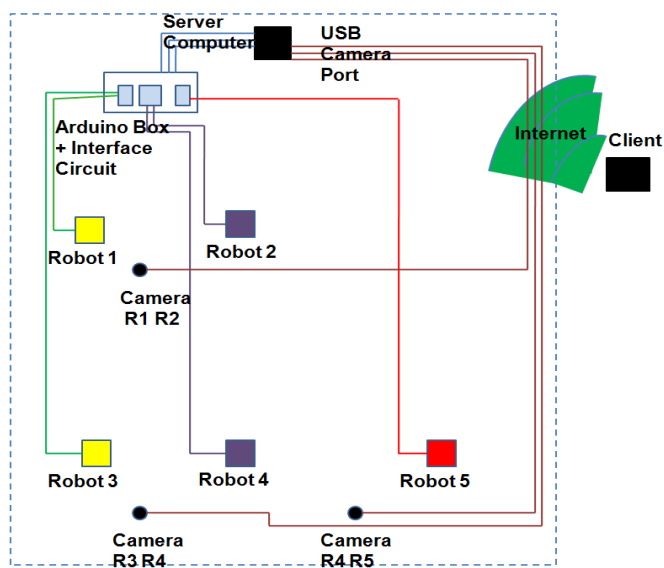


Fig. 10. Floor plan for future development of the system at the IIUM Robotics laboratory.

It is suggested that the system could be expand to multiple robots as shown in Fig. 10. It will reduce cost and help both operator and programmer give a positive impact for the students in terms of demonstrating the online and offline programming in Industrial Robotics. Students will then know that there are many optional for control the robot wirelessly. Furthermore, instead of CN5 input port, there are many other ports such as CN6~7 (for output devices) and CN10 (for controlling the robots motor). By having access on these ports, calibration can be done from the external mode. Other than input channel CN5, which allows us to run the calibration during robot star-up operation, is the CN6-11 (CAL completion). This function enables the programmer to know that the calibration is completed. It is done when the CAL and operation ready start signal is high.

Another suggestion would be to implement the CALSET operation automatically. CALSET refers to the calibration of the positional relation between the robot main unit and controller. It is performed whenever a mechanical end is changed, motor is replaced or the position data of the encoder are lost due to a random encoder backup battery.

CALSET can be done either by the mechanical ends or by entering XY coordinates. Both calibration methods perform almost at the same accuracy but the mechanical end method is easier and faster. Mechanical end can be accomplish by, move the arm by hand until it touches the mechanical end, and then record the position. This requires more operating space to allow the robot to travel to its allowable ends. Therefore, it is suggest that we could put the displacement sensor at every joint of the robots and getting the feedback form it to know the position of the motors.

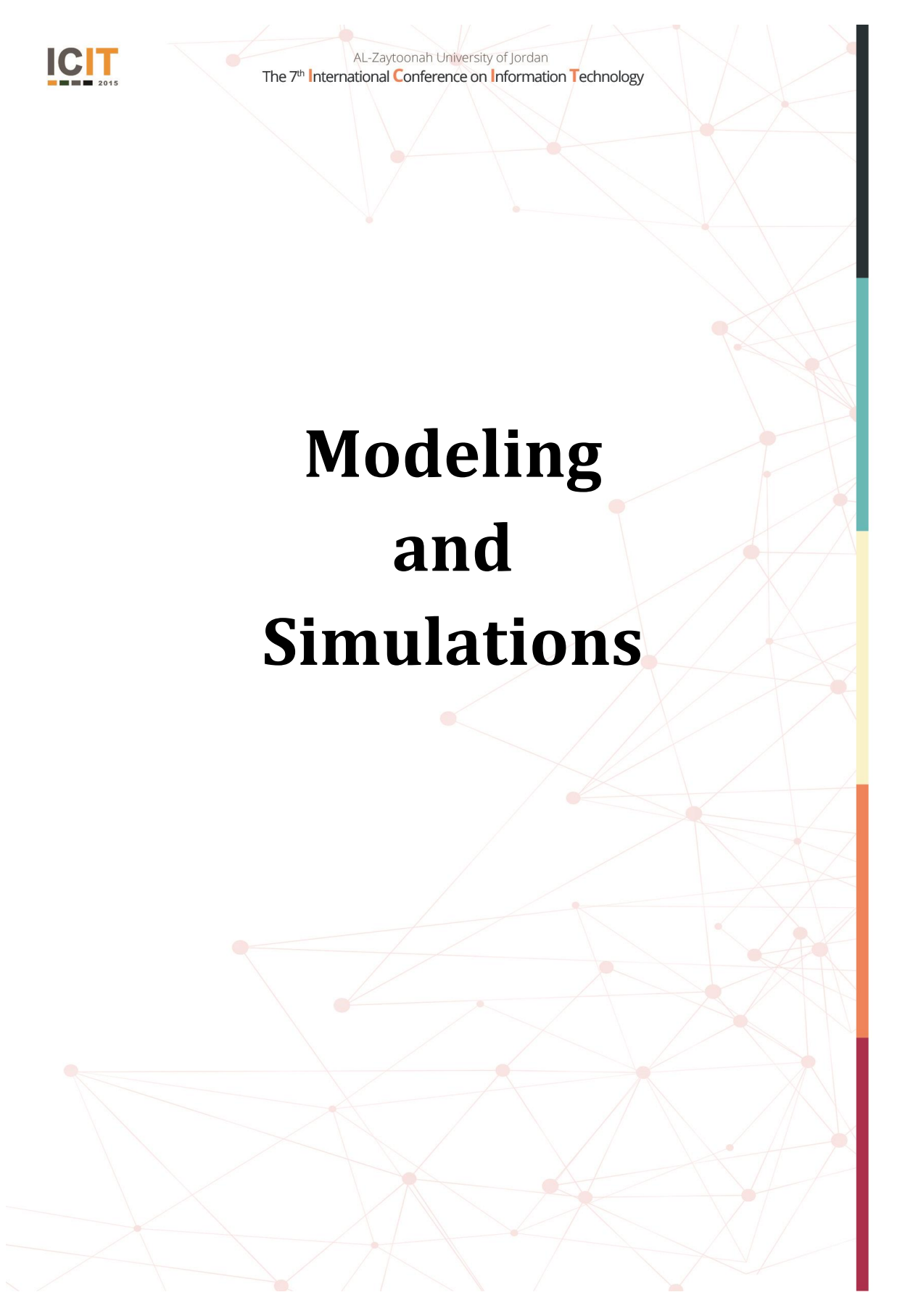
ACKNOWLEDGEMENT

We acknowledge the financial support of the Research Management Centre, IIUM, under Research Endowment Grant (EDW B14-161-1046).

REFERENCES

- [1] F. Tobe, "Low-cost robots like Baxter, UR5 and UR10 successfully entering small and medium enterprises (SMEs)," *Article, Robohub*, 2013. (<http://robohub.org/rethink-robotics-baxter-and-universal-robots-ur5-and-ur10-succeeding/>)
- [2] R. Marin, G. Leon, R. Wirz, J. Sales, J.M. Clever, P.J. Sanz and J. Fernandez, "Remote programming of network robot within the UJI industrial robotics telelaboratory," *IEEE Transactions On Industrial Electronics*, Vol. 56, No. 12, 2009, pp. 4806-4816.
- [3] A. Rosado-Muñoz, J. Muñoz-Marí, J. V. Francés-Villora and M. Bataller-Mompeán, "Remote Laboratory for Industrial Robot," *Transaction on Control and Mechanical Systems*, Vol. 2, No. 2, 2013, pp. 77-82.
- [4] L. Wang, A. Mohammed and M. Onori, "Remote robotic assembly guided by 3D models linking to a real robot," *CIRP Annals - Manufacturing Technology*, Elsevier. vol. 63, Issue 1, 2014, pp. 1-4.
- [5] A. Turan, S. Bogosyan & M. Gokasan, "Development of a client-server communication method for Matlab/Simulink based remote robotics experiments," *IEEE ISIE*, Montreal, Quebec, Canada, 2006.
- [6] C.S. Tzafestas, M. Alifragis, N. Palaiologou, S.C.A. Thomopoulos, M. Brahman, and A.E. Exarchou, "Development and experimental evaluation of remote laboratory platform for teaching robot manipulator programming" *Proceedings, Int. Conference on Engineering Education*, Florida, USA, 2004.

Modeling and Simulations



Multicore RISC Processor Implementation by VHDL for Educational Purposes

Safaa S. Omran and Ali J. Ibada

Department of computer engineering techniques
College of Electrical and Electronic Engineering Techniques
Baghdad, Iraq
omran_safaa@ymail.com , ali.alshukri@yahoo.com

Abstract— With trends computer manufacturers to build computers that have Multicore processors, it becomes necessary to study the hardware architecture of this processor and the way of manage data between Cores. All the previous researches were designing single cycle processors or pipeline processors by FPGA (Field Programmable Gate Array). This is a first research work on parallel processing to design and implement a Multicore processor by FPGA. In this work Multicore processor has two Cores and each Core consists of 5-stage pipeline MIPS (Microprocessor without Interlocked Pipeline Stages) RISC (Reduced Instruction Set Computer) processor. Separated data cache and instruction cache were added to each Core. MESI (Modified, Exclusive, Shared and Invalid) protocol is used to manage cache coherence and memory coherence which support Write-back policy where replacement algorithm is not needed. Many programs are tested on this design and the correct results were obtained. The VHDL (Very high speed integrated circuit Hardware Description Language) of the complete Multicore processor is implemented by using (Xilinx ISE Design Suite 13.4) Software and configured on FPGA Spartan-3AN starter kit and results from the kit were obtained.

Keywords— Multicore; MIPS; RISC; MESI protocol; VHDL; FPGA.

I. INTRODUCTION

Computer pioneers correctly predicted that programmers would want unlimited amounts of fast memory. An economical solution to that desire is a memory hierarchy, which takes advantage of locality and trade-offs in the cost performance of memory technologies. The principle of locality says that most programs do not access all code or data uniformly. Locality occurs in time (temporal locality) and in space (spatial locality) [1]. The different levels form what is commonly termed the memory hierarchy is a tiered description of how the different levels compare to and interact with each other. The different levels of the memory hierarchy are managed by different parts of the system [2]. On modern architectures a main memory access may take hundreds of cycles, so there is a real danger that a processor may spend much of its time just waiting for the memory to respond for requests. To alleviate this problem one or more caches are logically situated between the processor and the memory [3].

To get continuing performance gains of Multicore processor, it is requisite to use parallel software. Most parallel software relies on the shared-memory programming model in which all processors access the same physical address space, this cause cache coherency problem. To address the cache coherency problem, there are many protocols to deal with this [4]. In this paper MESI protocol is used.

Many previous researches have designed single Core (single cycle or pipeline processor) that can execution some instruction of MIPS processor [5-10]. In this work all

instructions are designed with extra (*hlt*) instruction that could be used to stop program execution.

VHDL is a VHSIC Hardware Description Language. VHSIC is an abbreviation for Very High Speed Integrated Circuit. It describes the behavior of an electronic circuit or system, such as ASICs (Application Specific Integrated Circuit) and FPGAs as well as conventional digital circuits. A fundamental motivation to use VHDL is that VHDL is a standard, technology/vendor independent language, and is therefore portable and reusable [11]. VHDL has Feature to allow the synthesis of a circuit or system in a programmable device. This paper studies the designing and prototyping of a complete design of Multicore MPIS RISC processor in VHDL. FPGA is a digital integrated circuit that contains configurable (programmable) blocks of logic along with configurable interconnects between these blocks. Design engineers can program such devices to perform a tremendous variety of tasks [12].

II. CACHE MEMORY PRINCIPLES AND DESIGN ELEMENTS

The cache contains a copy of portions of main memory [13]. When the processor attempts to read a word of memory, a check is made to determine if the word is in the cache. If so, the word is delivered to the processor. If not, a block of main memory, consisting of some fixed number of words, is read into the cache and then the word is delivered to the processor [14]. Each Core in the processor has its own cache and the cache lies on the same chip of the processor as shown in Figure 1. The cache has the following design choices:

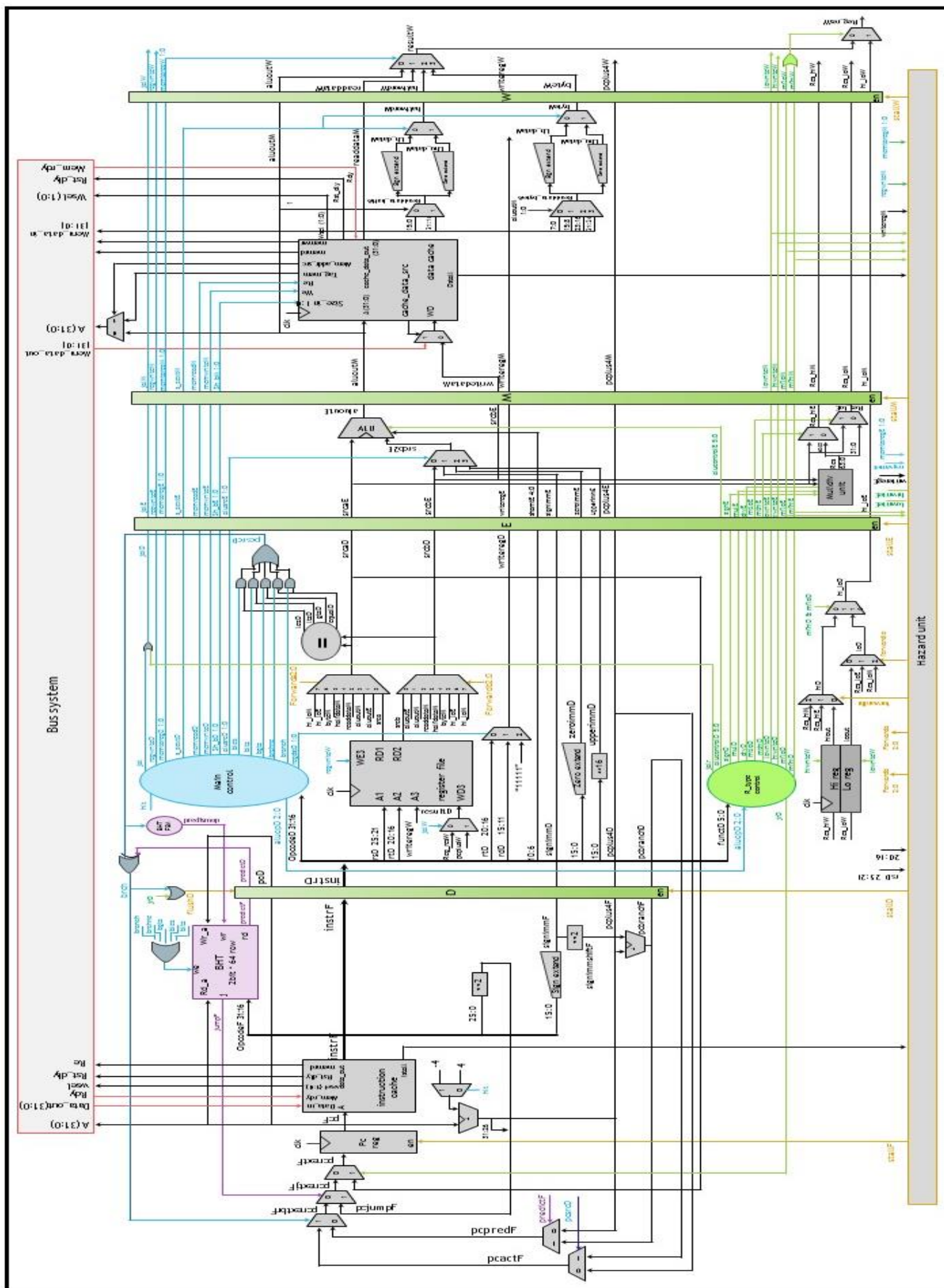


Figure 1 Complete design of one Core with cache memories

A. Number of Caches

The cache has been spliced into instruction cache and data cache to avoid structural hazards. These two caches both exist in the CPU (Central Processing Unit), typically as two level one (L1) caches. When the processor attempts to fetch an instruction from main memory, it first consults the instruction L1 cache, and when the processor attempts to fetch data from main memory, it first consults the data L1 cache.

B. Cache Addresses

The cache memory directly receives physical addresses instead of virtual addresses from the MIPS processor, for this reason there is no need to include a Memory Management Unit (MMU) in this design.

C. Mapping Function

Direct mapping is used in this design to map each block of main memory into only one possible cache line. Direct mapping is picked because there are fewer cache lines than main memory blocks.

D. Write policy

Write policy is needed for data cache only, because the processor will not update the program instruction. Write back policy is used in this design to minimize memory writes, where a copy of the data is written to data cache by the processor and not to main memory. When new data is written to cache, a MESI state is change to M (Modified) bits associated with the line is set. Then, when a block is replaced, it is written back to main memory and MESI state get the new line state.

E. Cache Size

For this design, each Core has cache size of 64 bytes, organized as 4 lines, each line has 4 words, and each word is 4 bytes in length. Therefore, main memory address is organized as shown in Figure 2.

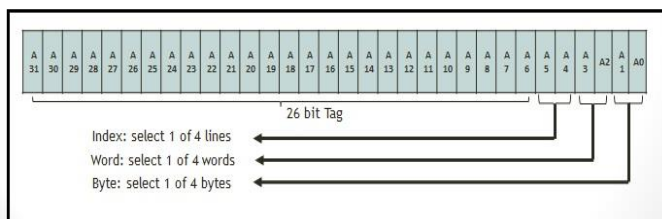


Figure 2 Address bit field format

III. CACHE COHERENCY MECHANISMS

To implement cache coherency protocols in the Multicore system and managed the consistency of memory, cache coherency mechanism is used. In this work Snooping Based Coherency is used as a cache coherency mechanism. It allows each cache to monitor the address lines so that to gain access to main memory which they have cached. Any activity on cache line will trigger message, which will be broadcasted to all the caches to update the cache line with the activity.

IV. CACHE COHERENCE PROTOCOL

In Multicore systems, coherence must occur inside each core and among cores through bus system. For this design MESI protocol is chosen. It is one of the mostly used cache coherency protocol. Any cache line can be in one of 4 states (2 bits):

- 1) Modified (00): cache line has been modified from the value in the main memory.
- 2) Exclusive (01): cache line is the same as main memory and is the only cached copy.
- 3) Shared (10): Same as main memory but copies exist in other caches.
- 4) Invalid (11): Line data is not valid (as in simple cache). So it should not be used.

A state transition diagram in Figure 3 shows what happens to a cache line in a processor as a result of memory accesses made by that processor (read hit/miss, write hit/miss). Memory accesses made by other processors that result in bus transactions observed by this snoopy cache (Mem RD, Mem WR, Invalidate) as shown in Figure 4.

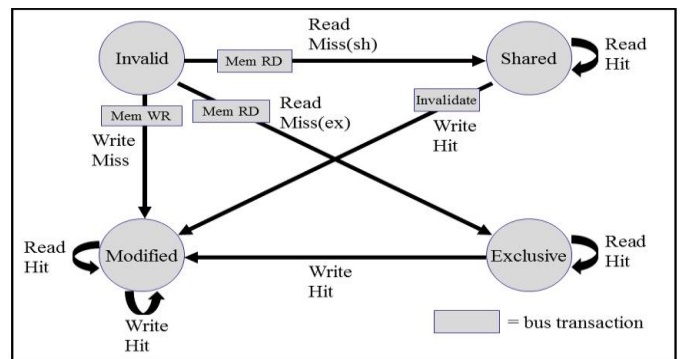


Figure 3 MESI – locally initiated accesses

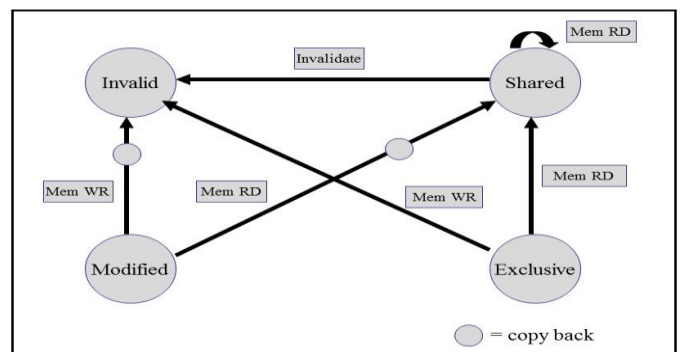


Figure 4 MESI – remotely initiated accesses

V. CACHE CONTROLLER

Cache controller is used to regulate cache memory. When Core wants to access memory location for read or write, it is first send address to its cache controller which decides this address is exists in tag cache or not. If it is, then no memory access is needed, the data is provided to Core directly from its cache; if not, then the cache controller fetches several words from main memory consecutively to fill the corresponding line

in the cache. Data cache controller use two bits for MESI protocol while instruction cache controller use one bit (valid or not valid only), because instruction cache does not make any change to instruction program. Also data cache controller has extra control signals to manage write-back that requested from bus system when a Core need to access a data that modified in another Core's cache, however instruction cache controller does not have this signals because there is no write-back in instruction cache. The cache controller consists of:

1) *Finite State Machine (FSM)*: the FSM of data cache differs from that of instruction cache because data is accessed for read or write while instructions are executed without modification. FSM of data cache is shown in Figure 5.

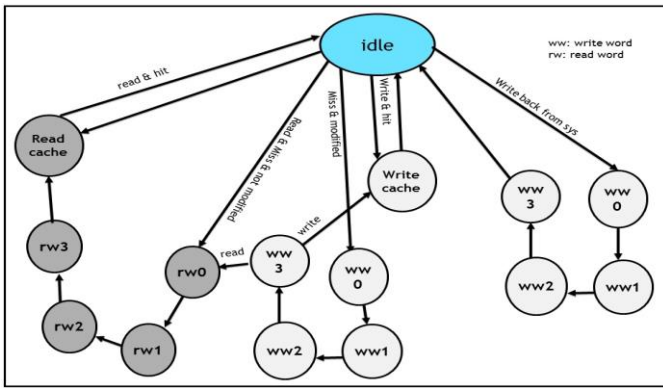


Figure 5 Data cache FSM

Write back from system has the priority to execute if read or write happen at the same time with system write back. Table 1 explains FSM work, and table 2 describes the function of FSM.

TABLE I. DATA CACHE FSM TRUTH TABLE

state	inputs				outputs										
	Hit	read	write	wb_in	MESI	stall	cachewr	cachedrd	memrd	memwr	Cache data src	Mem addr_src	Rst_dly	wsel	wb_done_out
idle (st0)	x	0	0	0	xx	1	0	0	0	0	x	x	1	00	1
Write cache (st1)	x	0	1	0	/=00	1	1	0	0	0	0	x	1	00	1
Read cache (st2)	1	1	0	0	xx	1	0	1	0	0	x	x	1	00	1
WW 0 (st3)	0	1 or 1	0	0	00	0	0	1	0	1	x	1	0	00	1
WW 1 (st4)	0	1 or 1	0	0	00	0	0	1	0	1	x	1	0	01	1
WW 2 (st5)	0	1 or 1	0	0	00	0	0	1	0	1	x	1	0	10	1
WW 3 (st6)	0	1 or 1	0	0	00	0	0	1	0	1	x	1	0	11	1
RW 0 (st7)	0	1	0	0	/=00	0	1	0	1	0	1	0	0	00	1
RW 1 (st8)	0	1	0	0	/=00	0	1	0	1	0	1	0	0	01	1

RW 2 (st9)	0	1	0	0	/=00	0	1	0	1	0	1	0	0	10	1
RW 3 (st10)	0	1	0	0	/=00	0	1	0	1	0	1	0	0	11	1
WW 0 (st11)	x	x	x	1	xx	0	0	1	0	1	x	1	0	00	0
WW 1 (st12)	x	x	x	1	xx	0	0	1	0	1	x	1	0	01	0
WW 2 (st13)	x	x	x	1	xx	0	0	1	0	1	x	1	0	10	0
WW 3 (st14)	x	x	x	1	xx	0	0	1	0	1	x	1	0	11	1

TABLE II. FUNCTION OF FSM SIGNALS

Signal name	Signal value	Signal effect
stall	0	Main memory is accessed and the whole pipeline is stalled.
	1	Cache memory is accessed and the pipelined registers are captured on the next falling edge.
cachewr	0	None
	1	When cache hit occurs, data supplied by the processor is written into cache memory.
cachedrd	0	None
	1	When cache hit occurs, data is supplied to the processor from cache memory.
memrd	0	None
	1	When cache miss occurs, data is supplied to the cache memory from main memory.
memwr	0	None
	1	When cache miss occurs and dirty bit is set, data block which is supplied by cache memory is written into main memory.
Cache_data_src	0	The value fed to the cache_data_in input of cache memory comes from the processor.
	1	The value fed to the cache_data_in input of cache memory comes from main memory.
Mem_addr_src	0	The address fed to the amem input of main memory comes from the processor.
	1	The address fed to the amem input of main memory equals to (tag & I & 0).
Rst_dly	0	The address fed to the amem input of main memory equals to (tag & I & 0).
	1	There is no main memory activity.
wsel	00	The first (least significant) word of memory block is selected.
	01	The second word of memory block is selected.
	10	The third word of memory block is selected.
	11	The fourth (most significant) word of memory block is selected.
wb_done_out	0	Cache controller is responding to write back request from bus system and Core is stall.
	1	Write back is done by cache controller and Core work properly.

FSM of instruction cache is part of FSM data cache that does not contain the states performing write actions and write back system. Figure 6 shows instruction cache FSM, and table 3 explains its work.

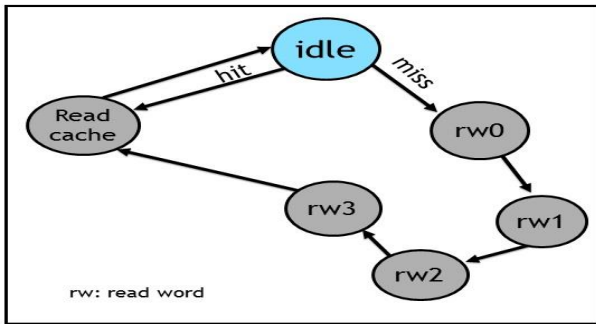


Figure 6 Instruction cache FSM

TABLE III. INSTRUCTION CACHE FSM TRUTH TABLE

state	inputs		outputs					
	Hit	Mem_rdy	stall	cachewr	cache rd	memrd	Rst_dly	wsel
idle (st0)	x	x	1	0	0	0	1	00
Read cache (st1)	1	1	1	0	1	0	1	00
Rw 0 (st2)	0	1	0	1	0	1	0	00
Rw 1 (st3)	0	0	0	1	0	1	0	01
Rw 2 (st4)	0	0	0	1	0	1	0	10
Rw3 (st5)	0	0	0	1	0	1	0	11

2) *Tag cache*: data tag cache contains 26 tag bits, 2 bits for MESI protocol for each data cache line. Tag bits are used for holding the 26 most significant bits of the address being accessed. MESI bits are reset when the machine restarts. Instruction tag cache is similar to data tag but does not have 2 bits for MESI protocol, instead it has 1 bit to indicate the line valid or not (valid bit).

VI. COMPLETE CACHE DESIGN AND MEMORY SYSTEM

For each Core, data cache controller is combined with its data cache as shown in Figure 7, while Figure 8 shows instruction cache controller that is combined with instruction cache.

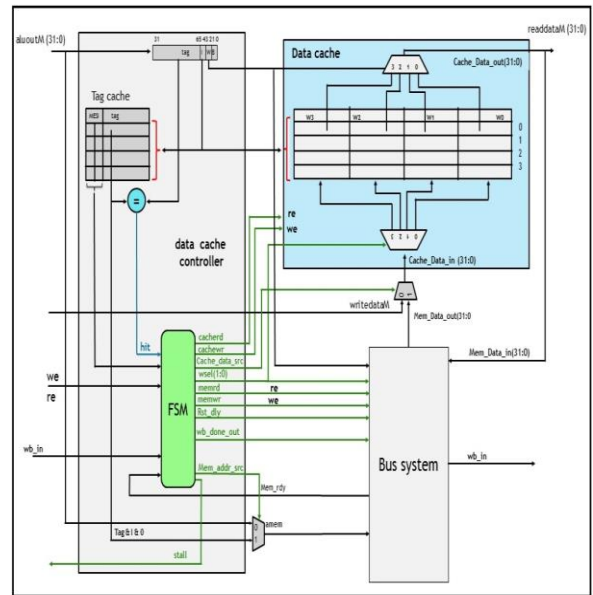


Figure 7 Complete design of data cache

Both caches access the main memory that consists of 1 kilobyte, arranged as 2 segments each one has 512 bytes; one segment for data and the other for instruction, each segment has 32 blocks, each block consists of 4 words and each word contains 4 bytes. Figure 9 shows main memory.

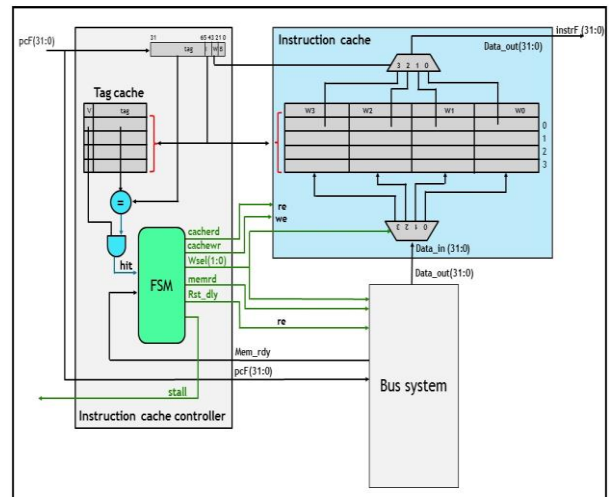


Figure 8 Complete design of instruction cache

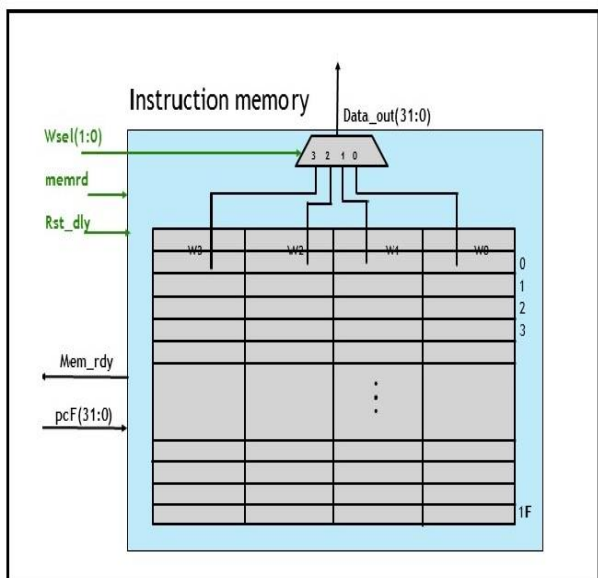


Figure 9 Main memory

VII. VHDL TOP-LEVEL IMPLEMENTATION

Top level of Multicore processor connects two Cores to data and instruction memories through bus system as shown in Figure 10. Later a test bench is written and used to execute a program.

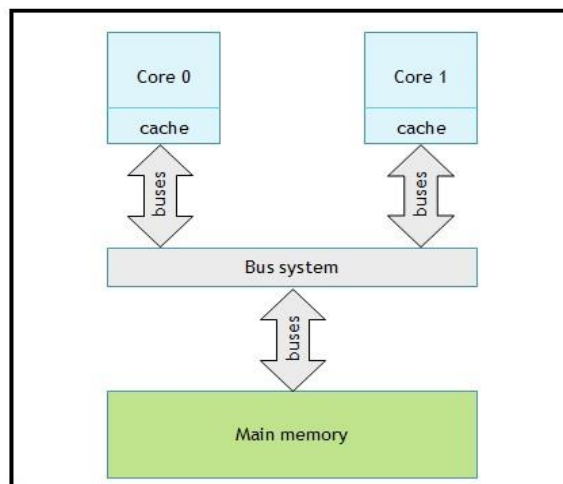


Figure 10 Multicore processor system

VIII. RESULTS

The test program shown in Figure 11 is stored in main memory. This program can be executed as a parallel code to get profit of Multicore system. This program used to find the summation of numbers (1 – 10)_h plus factorial of number 7. The results should be (00001438)_h stored in memory location (40)_h and (00000000)_h stored in memory location (44)_h.

	Assembly	address	discretion	machine
	addi \$t0,\$0,10	0	\$t0 = 10h	20080010
	addi \$t6,\$0,0	4	\$t6 = 0h	200E0000
loop:	add \$t6,\$t6,\$t0	8	\$t6 = \$t6 + \$t0	01C87020
	subi \$t0,\$t0,1	c	\$t0 = \$t0 - 1	2108FFFF
	bne \$t0,\$0,loop	10	if (\$t0 != \$zero)	1408FFFD
	goto loop		goto loop	
	sw \$t6,40(\$0)	14	mem[\$zero + 64] = \$t6	AC0E0040
	addi \$t3,\$0,7	18	\$t3 = 7	200B0007
	addi \$a0,\$0,1	1c	\$a0 = 1	20040001
loop2:	mult \$t3,\$a0	20	\$hi , low = (\$t3 * \$a0)	01640018
	mflo \$a0	24	\$a0 = \$lo	00002012
	mfhi \$a1	28	\$a1 = \$hi	00002810
	subi \$t3,\$t3,1	2c	\$t3 = \$t3 - 1	216BFFFF
	bne \$t3,\$0,loop2	30	if (\$t3 != \$zero)	140BFFFF
	goto loop2		goto loop2	
	lw \$v0,40(\$0)	34	\$v0 = mem[\$zero + 64]	8C020040
	add \$a0,\$a0,\$v0	38	\$a0 = \$a0 + \$v0	00822020
	sw \$a0,40(\$0)	3c	mem[\$zero + 64] = \$a0	AC040040
	sw \$a1,44(\$0)	40	mem[\$zero + 68] = \$a1	AC050044
	hlt	44	stop program execution	F0000000

Figure 11 Top level test program

This program has been executed as a parallel code in Multicore processor. By using VHDL testbench, the right results have been gotten as shown in Figure 12 which indicates the correctness of the design. When memwrite signal is 1, the results are stored in data memory.

The program shown in Figure 11 is executed in single core system and Multicore system to make a comparison in terms of

performance and speedup between single Core processor and Multicore processor as shown in table 4. The CPI (Clock Per Instruction) metric is calculated by using equation:

$$\text{Program Execution time} = \frac{\text{Instruction count} \times \text{CPI} \times \text{Clock period}}{\dots\dots\dots} \quad (1)$$

TABLE IV. PERFORMANCE COMPARISON BETWEEN SINGLE CORE AND MULTICORE PROCESSORS

Processor	Instruction count	Program execution time	No of clock cycles	Clock period	CPI	Speedup
Single Core	92	1255	125.5	10 ns	1.36	1
Multicore	92	865	86.5	10 ns	0.94	1.45

This design is configured on Xilinx Spartan-3AN starter kit FPGA. To show all results, VGA (Video Graphic Array) screen is interfaced with FPGA via a standard high-density HD-DB15 female connector VGA display port and driving the VGA monitor in 640 by 480 mode. Figure 13 shows results of test program on VGA screen. The left column is an assembly test program machine code with its locations in instruction memory that would the processor fetches it to be execution. Drawing in the center is illustration that the Processor is connected to data and instruction memories via buses. The Right column represents the data memory that would the

processor uses it to store or load data, results of test program are shown in data column with its locations.

IX. CONCLUSIONS

VHDL design of Multicore RISC processor has been implemented for whole instructions which consist of 49 instructions. Also hlt instruction was added to stop program execution. Each Core was Pipelined to five stages. MESI protocol was used to deal with data coherence which represents the main problem of Multicore system. On chip cache system was added for each Core. Cache system used direct mapping function, write back policy. The cache system consists of two separated caches; one for data and one for instruction. After all system design was completed, various programs simulated and results were obtained. It is meaning that design work properly. The Xilinx ISE Design Suite 13.4 program is used for design synthesis while the Xilinx ISim simulator program is used to simulate this design which is then configured on a Xilinx Spartan-3AN FPGA starter kit and results from kit were obtained.

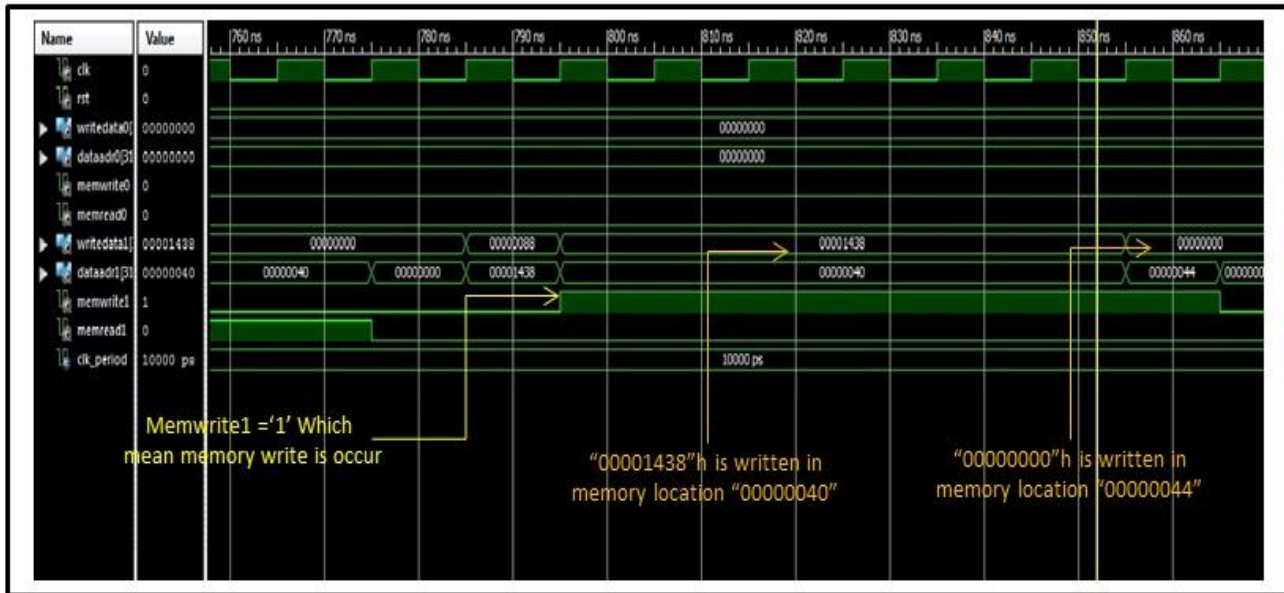


Figure 12 Simulation waveform of test program

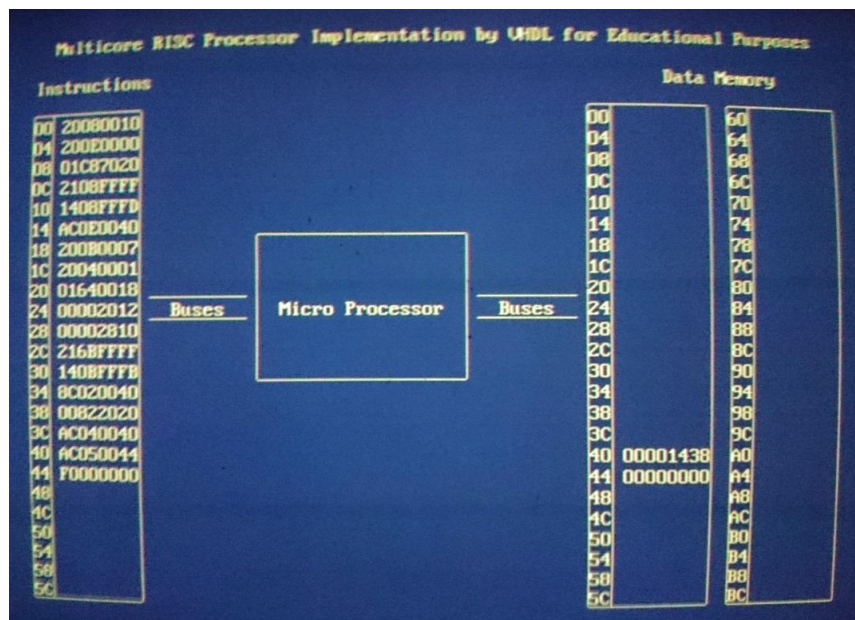


Figure 13 Results of test program on VGA screen

REFERENCES

- [1] J. L. Hennessy and D. A. Patterson, "Computer Architecture: A Quantitative Approach", 5th ed., San Francisco, USA: Morgan Kaufmann, 2012.
- [2] D. Page, "A Practical Introduction to Computer Architecture", London, UK: Springer-Verlag, 2009.
- [3] M. Herlihy and N. Shavit, "The Art of Multiprocessor Programming", Burlington, USA, Morgan Kaufmann, 2008.
- [4] S. Dey, and M. S. Nair, "Design and Implementation of a Simple Cache Simulator in Java to Investigate MESI and MOESI Coherency Protocols", International Journal of Computer Applications, Vol. 87 – No.11, 0975 – 8887, 2014.
- [5] M. B. I. Reaz, Sh. Islam and M. S. Sulaiman, "A Single Clock Cycle MIPS RISC Processor Design using VHDL", IEEE International Conference on Semiconductor Electronics (ICSE2002), Penang, Malaysia, PP. 126 – 129, DEC. 2002.
- [6] S. P. Katke and G. P. Jain, "Design and Implementation of 5 Stages Pipelined Architecture in 32 Bit RISC Processor", International Journal of Emerging Technology and Advanced Engineering, vol. 2, no. 4, PP. 340-346, Apr. 2012.
- [7] V. Robio, "A FPGA Implementation of A MIPS RISC Processor for Computer Architecture Education", MSc. thesis, New Mexico State University, Las Cruces, New Mexico, America, 2004.
- [8] B. valli, A. U. Kumar and B. V. Bhaskar, "FPGA Implementation and Functional Verification of a Pipelined MIPS Processor", International Journal Of Computational Engineering Research, Vol. 2, No. 5, PP. 1559-1561, Sep. 2012.
- [9] I. Anthony, "VHDL Implementation of Pipelined DLX Microprocessor", MSc. Thesis, University Teknologi Malaysia (UTM), Malaysia, 2008.
- [10] H. Mahmood and S. omran, "Pipelined MIPS Processor with Cache Controller using VHDL Implementation for Educational Purposes", International Conference on Electrical Communication, Computer, Power, and Control Engineering ICECCPCE1, Mosul, Iraq, 2013.
- [11] Pedroni V., "circuit design with VHDL", MIT Press, London, England, 2004.
- [12] C. Maxfeild, The Design Warrior's Guide to FPGAs: Devices, Tools and Flows, Burlington, USA: Elsevier, 2004.
- [13] M. Abd-El-Barr and H. El-Rewini, Fundamentals of Computer Organization and Architecture, New Jersey, USA: John Wiley & Sons, 2005.
- [14] W. Stallings, Computer Organization and Architecture: Designing for Performance, 8th ed., New Jersey, USA: Pearson Education, 2010.

Numerical Simulation of Axial Coolant Flow in Rod Bundles of a Nuclear Reactor

Y. Rihan

Nuclear Fuel Tech. Dep.
Atomic Energy Authority, Hot Lab. Center,
Egypt.
yarihan159@yahoo.com

I. Salama

Nuclear Fuel Tech. Dep.
Atomic Energy Authority, Hot Lab. Center,
Egypt.

Abstract— Numerical simulation of coolant flow inside the rod bundle of a nuclear reactor is of great engineering interest. In the design of innovative core solutions, such as high conversion tight lattice cores for Light Water Reactors (LWR), as flow distribution cannot be calculated with exact analytical methods, numerical modeling plays a vital role. A computational fluid dynamics (CFD) methodology is proposed to investigate the thermal–hydraulic characteristics in a rod bundle. Using a three dimensional numerical solution, the characteristics of an isotropic k-epsilon turbulence model for use in modeling turbulent interchange mixing within rod arrays was investigated. The model used to predict the radial component of turbulent eddy viscosity and wall shear. Existing data of Nusselt number distributions in the axial direction obtained by different authors have been employed to validate the CFD model.

Keywords— Coolant flow; Rod bundles; Nuclear; Turbulent; Modeling

I. INTRODUCTION

Understanding the physical behavior of the coolant as it flows through the fuel bundles is of interest to those analyzing reactor operation and safety. An important aspect of understanding this behavior is the mixing of coolant momentum and heat. Without adequate coolant mixing models, the heat removal capabilities and safety margins of the reactor cannot be accurately established nor predicted. The nuclear fuel assemblies of Pressurized Water Reactors (PWR) consist of rod bundles arranged in a square configuration. The constant distance between the rods is maintained by spacer grids placed along the length of the bundle. The coolant flows mainly axially in the subchannels formed between the rods. Most spacer grids are designed with mixing vanes which cause a cross and swirl flow between and within the subchannels, enhancing the local heat transfer performance in the grid vicinity.

Many nuclear subchannel analysis codes adopt the lumped parameter approach, where many empirical correlations are used to simplify the complex exchange phenomena between subchannels. Therefore, the prediction capability of a subchannel analysis code depends thoroughly on the pertinent usage of the models and correlations [1]. A subchannel is defined as the flow area bounded by a cluster of fuel rods. In

most rod bundle configurations, two basic subchannel geometries are found: square and triangular as shown in Fig. 1.

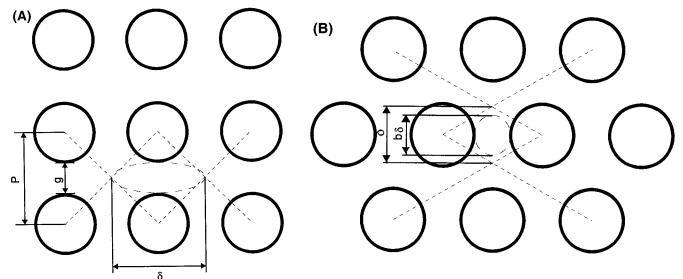


Fig. 1. Schematic view of rod bundle geometry of (a) square array; (b) triangular array

There have been several studies on flow mixing and heat transfer enhancement caused by a coolant flow in rod bundle geometry. Amongst the many studies performed involving CFD (Computational Fluid Dynamic) simulations of rod bundles with spacer grids some of the most significant contributions are: Karoutas et al. [2] and Imaizumi et al. [3] that demonstrated the usefulness of single subchannel CFD methodologies coupled with experimental results from LDV (Laser Doppler Velocimetry) and pressure loss measurements on the development of fuel designs for PWR reactors. Navarro

et al. [4] used the $k-\varepsilon$ model that presents results of flow simulations performed with the CFD code in a PWR 5×5 rod bundle segment with a split-vane spacer grid. Holloway et al. [5] showed that there is a great variation of heat transfer distribution along a fuel rod due to the spacer grid type. Wu and Trupp [6] clearly demonstrated that flow conditions inside the fuel bundles are very different from those in the typical pipes. The near-wall turbulence anisotropy results in the formation of secondary vortices inside the channel, causing the coolant to spiral through the bundle. Liu et al. [7] presented the results of numerical issues such as mesh refinement, wall treatment and appropriate definition of boundary conditions, which exert great influence on the results of a CFD simulation.

Recently, simulation studies were performed to investigate the thermal-hydraulic phenomena within a rod bundle [8-15]. Ga'bor [16] demonstrated that the Reynolds stress model (RSM) could be a good candidate for the accurate modeling of rod bundles. Baglietto and Ninokata [17] show that a quadratic $k-\varepsilon$ model with adjusted coefficients can reproduce the wall shear stress and the velocity distributions a fully developed flow in a triangular lattice bundle. A CFD model was developed by Lin et al. [18] to investigate the flow characteristics in the rod bundle with the different pressure-strain models in RSM, including linear pressure-strain (LPS), Quadratic Pressure-Strain (QPS) and low-Re stress-omega (LROS) models using ANSYS FLUENT solver. Subchannel model was developed and the study of mesh sensitivity was performed initially. Most of previous CFD studies for rod bundles were focused on the hydraulic simulation. In addition, previous simulation works neglect the thickness of vane-pair spacer grids for modeling simplification. This simplification would increase about 15% flow area in the grid region and lower the flow velocity, which may result in different flow and heat transfer characteristics.

This paper introduces a mathematical model to understand the physical behavior of coolant mixing within a nuclear fuel bundle and assess its applicability to this study.

II. THE MATHEMATICAL MODEL

A mathematical model was developed in this paper to investigate the flow characteristics in rod bundles subchannels. This simulation mathematical model includes the continuity equation, momentum equation, energy equation, and $k-\varepsilon$ turbulence model.

Continuity equation:

$$\frac{\partial(u_i)}{\partial x_i} = 0 \quad (1)$$

Momentum equation:

$$\rho \frac{\partial u_i u_j}{x_j} = -\frac{\partial P}{\partial x_i} + \frac{\partial}{\partial x_j} \left[\mu \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) - \rho u'_i u'_j \right] - \rho g_i \quad (2)$$

Energy equation:

$$\frac{\partial}{\partial x_i} (u_i (\rho E + P)) = \frac{\partial}{\partial x_i} \left(K_{eff} \frac{\partial T}{\partial x_i} + u_j (\tau_{ij})_{eff} \right)$$

(3)

$k-\varepsilon$ turbulence model:

The transport of turbulent kinetic energy per unit mass in high Reynolds number form can be provided by the following equation [19].

$$\rho \frac{\partial k}{\partial t} + \rho u_i \frac{\partial k}{\partial x_i} = \frac{\partial}{\partial x_i} \left[\left(\mu + \frac{\mu_t}{\sigma_k} \right) \frac{\partial k}{\partial x_i} \right] + P_k - \rho \varepsilon + \frac{\rho \mu_t}{Pr_t} \beta \left(g_i \frac{\partial T}{\partial x_i} \right) \quad (4)$$

Where P_k is the volumetric production of k and can be expressed as follow:

$$P_k = \left[\mu_t \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) \right] \frac{\partial u_i}{\partial x_j} \quad (5)$$

The high Reynolds number form of the transport equation for the turbulence dissipation rate is given by the following.

$$\rho \frac{\partial \varepsilon}{\partial t} + \rho u_i \frac{\partial \varepsilon}{\partial x_i} = \frac{\partial}{\partial x_i} \left[\left(\mu + \frac{\mu_t}{\sigma_\varepsilon} \right) \frac{\partial \varepsilon}{\partial x_i} \right] + \frac{\varepsilon}{k} (c_{\varepsilon 1} P_k + c_{\varepsilon 2} \rho \varepsilon) \quad (6)$$

The empirical constants used in the k and ε equations are summarized in Table I.

TABLE I. $k-\varepsilon$ MODEL CONSTANTS

c_μ	$c_{\varepsilon 1}$	$c_{\varepsilon 2}$	σ_k	σ_ε	Pr_t
0.09	1.44	1.92	1.0	1.3	0.9

A. Boundary conditions

The wall boundary condition must account for the influence of the three layers (laminar, buffer, and logarithmic) on momentum transport to the wall. This is achieved by using a law of the wall for the momentum equation's wall boundary condition.

$$U^+ = \frac{u}{u_\tau} = \frac{\ln y^+}{0.41} + 5.2 \quad (7)$$

The dimensionless distance from the wall, y^+ , is defined in terms of the shear velocity, u_τ , and wall distance, y .

$$y^+ = \frac{\rho u_\tau y}{\mu} \quad (8)$$

where,

$$u_\tau = \sqrt{\tau_w / \rho} \quad (9)$$

A similar law of the wall is used as a boundary condition for the energy equation in the turbulent flow. For a specified wall temperature, T_w , the wall heat flux, q_w , is given by the following.

$$q_w = \frac{u_\tau}{T^+} (T_w - T)$$

$$(10)$$

The dimensionless temperature can be evaluated as following [20].

$$T^+ = \text{Pr} y^+ e^{-\Lambda} + (2.12 \ln y^+ + \beta^*) e^{-1/\Lambda} \quad (11)$$

$$\Lambda = \frac{0.01(\text{Pr} y^+)^4}{1 + 5 \text{Pr}^3 y^+} \quad (12)$$

$$\text{Pr} = \frac{\mu c_p}{k} \quad (13)$$

$$(14)$$

$$\beta^* = (3.85 \text{Pr}^{1/3} - 1.3)^2 + 2.12 \ln \text{Pr} \quad \text{The wall boundary}$$

condition for the k and ε equations is based upon assuming the production of turbulence equals dissipation, constant shear layer, and the velocity gradient normal to the wall is much greater than the gradient along the wall.

$$(15)$$

$$\varepsilon = \frac{u_\tau^3}{0.41 y}$$

The production of turbulent kinetic energy, P_k , can be calculated as follow:

$$P_k = \frac{\tau_w^2}{\mu} \frac{dU^+}{dy^+} \quad (16)$$

The inlet conditions for the k and ε equations utilize a specified turbulence intensity, T_u , and eddy length scale, L_ε .

$$k = \frac{3}{2} T_u^2 U_b^2 \quad (17)$$

$$\varepsilon = \frac{k^{2/3}}{L_\varepsilon} \quad (18)$$

The turbulence intensity is usually specified as 0.05, while the eddy length scale is specified as equal to a domain characteristic length, such as the radius for flow in a circular pipe.

The average dimensionless Nusselt number (Nu_{ave}) offers an insight on convective heat transfer that occurs at the surface of the rods. The average Nusselt number is defined by the following equation:

$$Nu_{avg} = \frac{h D_h}{k} \quad (19)$$

The local Nusselt number and normalized Nusselt number around the circumference of the rod bundle at each axial location are given by the expressions,

$$Nu = \frac{h(z) D_h}{k} \quad (20)$$

and

$$\frac{Nu(z) - Nu_{avg}(z)}{Nu_{avg}(z)} \quad (21)$$

The conservation and transport equations are solved on a discretized domain by integrating the differential equations over discrete control volumes. Flux terms are solved at integration points which ensure a strongly conservative solution. The weak coupling between pressure and velocity is treated by splitting the numerical evaluation of velocity into mass-flow and momentum-flow terms. This allows the use of collocated grids. A Gauss-Seidel solver is used which has the property that high-frequency error components are eliminated faster than low-frequency components.

III. RESULTS AND DISCUSSION

The measured data of Nu number for a rod bundle obtained by Holloway et al. [21] are used to validate the present model. Figure 2 shows the comparison of normalized Nu number along the axial location of the rod bundle between the measured data and the model predictions. The coolant fluid was the air in this experiment. The normalized Nu number is the average Nu number (Nu_{avg}) divided by $Nu_{avg,\infty}$. The predicted Nu_{avg} number is obtained by averaging the local Nu number around the azimuthal angle. As clearly revealed in Fig. 2, the predicted distribution of normalized Nu number agrees well with the measured data. The comparison reveals that k - ε turbulence model is suitable to be applied in simulating the flow and heat transfer in the rod bundles.

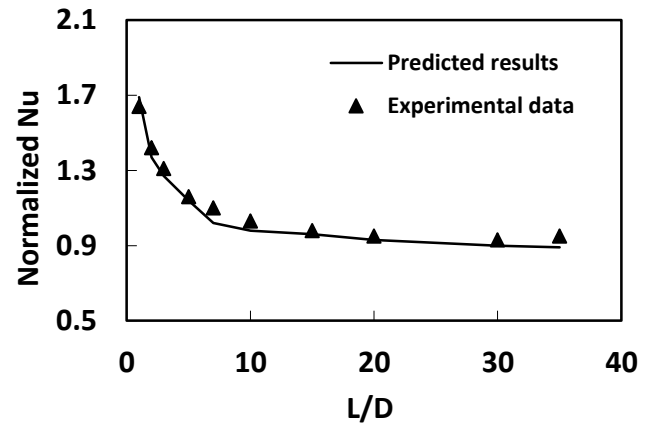


Fig. 2. Comparison between the predicted normalized Nu and the experimental data of Holloway et al. [21]

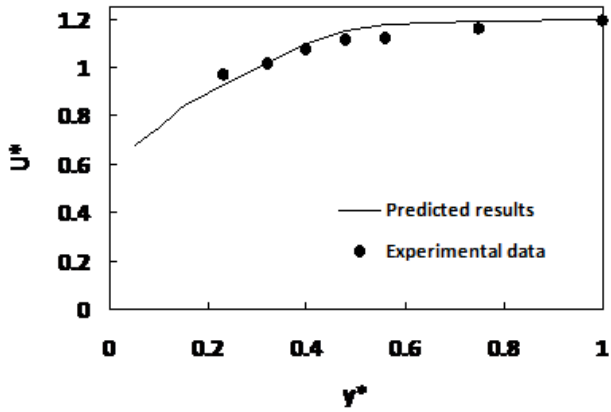


Fig. 3. Comparison between the predicted and the experimental data of Trupp et al. for normalized axial velocity [22]

A comparison of normalized axial velocity distributions (U^*) along the normalized distance from the wall ($y^* = y/L$) between the measurements data of Trupp et al. [22] and the predictions is shown in Fig. 3. As shown in this figure, the predicted distributions agree well with the measured data. This figure also reveals that the velocity distribution predicted by $k-\epsilon$ turbulence model is suitable to be applied in simulating the flow in the rod bundles.

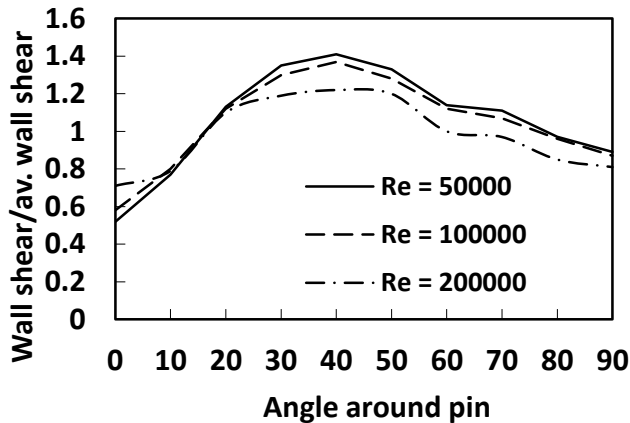


Fig. 4. The wall shear around rod, $\gamma_p = 1.1$

The mathematical model predicted variation in wall shear around the rod surface with a pitch-to-diameter ratio (γ_p) of 1.1 and different Reynolds numbers is shown in Fig. 4. The variation in wall shear near the gap is promotional to the variation in turbulence kinetic energy. The predicted maximum wall shear stress occurs approximately at angle 40 degrees from the gap and is constant with respect to Reynolds number. The effect of γ_p on the wall shear at constant Reynolds number is shown in Fig. 5. The wall shear distribution is more flat as γ_p increases and there is a definite movement of the location of peak wall shear towards the gap ($\theta = 0$). The wall shear is

proportional to the radial gradient of axial velocity, this leads to the movement of the position of maximum wall shear into the gap for larger values of γ_p .

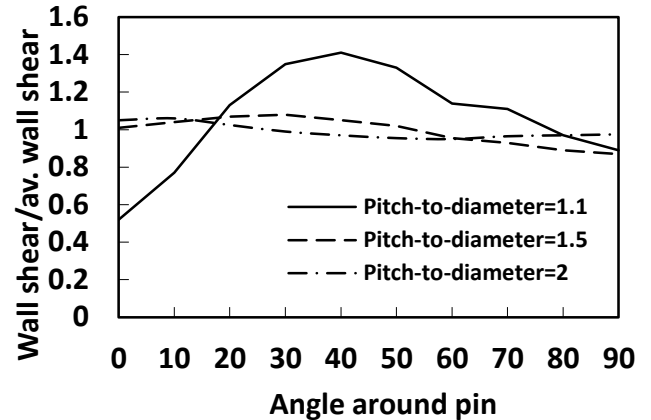


Fig. 5. The wall shear around rod, $Re = 50000$

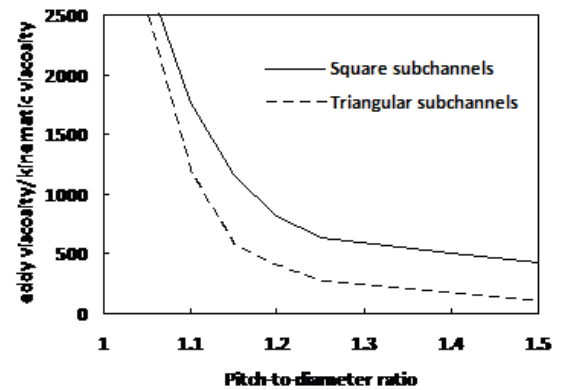


Fig. 6. Effect of pitch-to-diameter ratio on eddy viscosity, $Re = 100000$

The turbulent mixing increases with decreasing the gap size as shown in Fig. 6. The observed increase in mixing as γ_p decreases may be due to an increase in turbulence intensity in the gap. The predicted variation of turbulent eddy viscosity across the gap at different Re and for each of the pitch-to-diameter ratios are provided in Fig. 7 and Fig. 8.

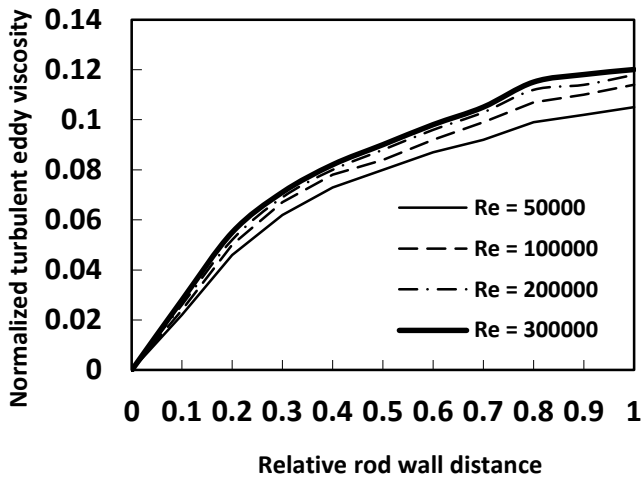


Fig. 7. Normalized eddy viscosity at different Re, $\gamma_p = 1.1$

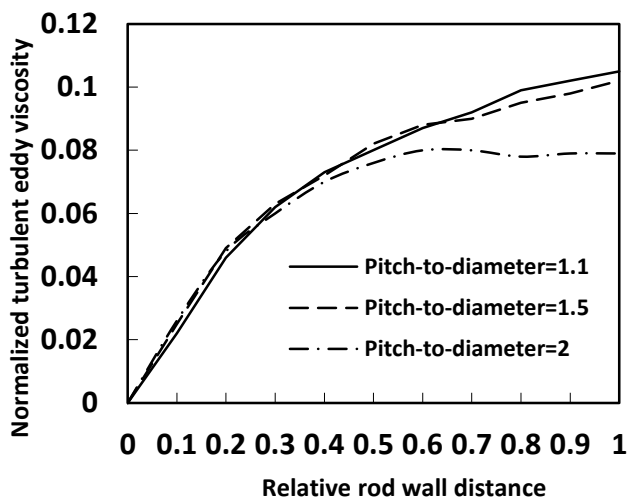


Fig. 8. Normalized eddy viscosity at different pitch-to-diameter ratios, Re = 50000

IV. CONCLUSIONS

A mathematical model of steady, three-dimensional turbulent fluid flow was presented in conjunction with a numerical solution procedure based upon finite volumes. Turbulence was modeled using an isotropic, $k-\varepsilon$ eddy viscosity model. Existing data of Nusselt number distributions in the axial direction obtained by different authors have been employed to validate the CFD model. Compared with the measured Nu distributions for different authors the present predicted results show good agreement. These comparisons reveal that the $k-\varepsilon$ turbulence model can be applied to reasonably simulate the flow and heat transfer behaviors for the rod bundle.

Nomenclature

c_p	heat capacity at constant temperature, J/kg.K
D_h	equilibrium diameter, m
E	modeling constant
k	turbulence kinetic energy, m^2/s^2
P	pressure, N/m ²
Pr_t	turbulent Prandtl number
T	temperature, K
u_i	velocity vector, m/s
u'_i	turbulent fluctuating quantity
U_b	axial bulk velocity, m/s
x_i	coordinate vector, m

Greek letter

β	fluid thermal expansion coefficient, K ⁻¹
ε	turbulence dissipation rate, m^2/s^3
μ	dynamic viscosity, kg/m.s
ρ	density, kg/m ³
τ_{ij}	Reynolds stress tensor, Pa
τ_w	wall shear, Pa

REFERENCES

- [1] S. Kim, and B. Chung, "A scale analysis of the turbulent mixing rate for various Prandtl number flow fields in rod bundles," *Nuclear Engineering and Design*, vol. 205, pp. 281–294, 2001.
- [2] Z. Karoutas, C. Gu, and B. Sholin, "3-D flow analyses for design of nuclear fuel spacer," In: *Proceedings of the 7th International Meeting on Nuclear Reactor Thermal-hydraulics NURETH-7*, New York, USA, 1995, pp. 3153-3174.
- [3] M. Imaizumi, T. Ichioka, M. Hoshi, H. Teshima, H. Kobayashi, and T. Yokoyama, "Development of CFD method to evaluate 3-D flow characteristics for PWR fuel assembly," In: *Transactions of the 13th International Conference on Structural Mechanics in Reactor Technology SMIRT 13*, Porto Alegre, Brazil, 1995, pp. 3-14.
- [4] M.A. Navarro, and A.C. Santos, "Evaluation of a numeric procedure for flow simulation of a 5x5 PWR rod bundle with a mixing vane spacer," *Prog. Nucl. Energy*, 2011.
- [5] M.V. Holloway, D.E. Beasley, and M.E. Conner, "Single-phase convective heat transfer in rod bundles," *Nucl. Eng. Des.*, vol. 238, pp. 848-858, 2008.
- [6] X. Wu, and A. C., "Trupp, Experimental study on the unusual turbulence intensity distribution in rod-to-wall gap regions," *Exp. Therm. Fluid Sci.*, vol. 6 (4), pp. 360-370, 1993.
- [7] C.C. Liu, Y. M. Ferng, and C.K. Shih, "CFD evaluation of turbulence models for flow simulation of the fuel rod bundle with a spacer assembly," *Applied Thermal Engineering*, vol. 40, pp. 389-396, 2012.
- [8] M.A. Navarro, and A. Santos, "Evaluation of a numeric procedure for flow simulation of a 5x5 PWR rod bundle with a mixing vane spacer," *Progress in Nuclear Energy*, vol. 53, pp. 1190-1196, 2011.
- [9] Y.S. Tseng, Y.M. Ferng, and C.H. Lin, "Investigating flow and heat transfer characteristics in a fuel bundle with split-vane pair grids by CFD methodology," *Annals of Nuclear Energy*, vol. 64, pp. 93–99, 2014.

- [10] X.J. Liu, T. Yang, and X. Cheng, "Thermal-hydraulic analysis of flow blockage in a supercritical water-cooled fuel bundle with sub-channel code," *Annals of Nuclear Energy*, vol. 59, pp. 194–203, 2013.
- [11] S. Chen, Y. Liu, T. Hibiki, M. Ishii, Y. Yoshida, I. Kinoshita, M. Murase, and K. Mishima, "One-dimensional drift-flux model for two-phase flow in pool rod bundle systems," *International Journal of Multiphase Flow*, vol. 40, pp. 166–177, 2012.
- [12] X. Li, and Y. Gao, "Methods of simulating large-scale rod bundle and application to a 17×17 fuel assembly with mixing vane spacer grid," *Nuclear Engineering and Design*, vol. 267, pp. 10–22, 2014.
- [13] A.K. Chauhan, B.V. Prasad, and B.S. Patnaik, "Thermal hydraulics of rod bundles: The effect of eccentricity," *Nuclear Engineering and Design*, vol. 263, pp. 218–240, 2013.
- [14] N.D. Patil, P.K. Das, S. Bhattacharyya, and S.K. Sahu, "An experimental assessment of cooling of a 54-rod bundle by in-bundle injection," *Nuclear Engineering and Design*, vol. 250, pp. 500–511, 2012.
- [15] X. Zhang, and S.D. Yu, "Large eddy simulation of turbulent flow surrounding two simulated CANDU fuel bundles," *Nuclear Engineering and Design*, vol. 241, pp. 3553–3572, 2011.
- [16] H. Ga'bor, "On turbulence models for rod bundle flow computations," *Annals of Nuclear Energy*, vol. 32, pp. 755–761, 2005.
- [17] E. Baglietto, and H. Ninokata, "A turbulence model study for simulating flow inside tight lattice rod bundles," *Nuclear Engineering and Design*, vol. 235, pp. 773–784, 2005.
- [18] C. Lin, C. Yen, and Y. Ferng, "CFD investigating the flow characteristics in a triangular-pitch rod bundle using Reynolds stress turbulence model," *Annals of Nuclear Energy*, vol. 65, pp. 357–364, 2014.
- [19] A.D. Young, "Boundary layers," *BSP Professional Books*, Oxford, 1989.
- [20] B.A. Kader, "Temperature and concentration profiles in fully turbulent boundary layers," *Int. J. Heat Mass Transfer*, vol. 29 (9), pp. 1541-1544, 1981.
- [21] M.V. Holloway, T.A. Conover, H.L. McClusky, and D.E. Beasley, "The effect of support grid design on azimuthal variation in heat transfer coefficient for rod bundles," *J. Heat Transfer*, vol. 127, pp. 598–605, 2005.
- [22] A.C. Trupp, and R.S. Azad, "The structure of turbulent flow in triangular array rod bundles," *Nuclear Engineering and Design*, vol. 32, pp. 47–84, 1975.

VeSimulator A Location-Based Vehicle Simulator Model for IoT Applications

Osama Oransa

Arab Academy for Science, Technology and Maritime Transport
Cairo, Egypt
osama_oransa@hotmail.com

Mostafa Abdel-Azim

College of Computing & Information Technology
Arab Academy for Science, Technology and Maritime Transport
Cairo, Egypt
melbakary@aast.edu

Abstract— Location-based services (LBS) are important aspects in today business models where a lot of use cases are built around the identification of user location such as the advertisements, asset tracking, geo-fence, and a lot of different things that utilize such location information. With the maturity of Global Positioning System (GPS) technology many different devices are shipped with built-in GPS unit, this enables the Internet of Things (IoT) application to utilize this location information and build many location-based business models including health, transportation, marketing and social services. Smart transportation is one of these important IoT applications in which LBS play a major role. In this research we built a location-based simulator model that can be used in developing internet of things applications. The research focused on developing a generic simulator model that can be customized according to the used application, so we can develop IoT location-based application without the need to have hardware components in early research phases. We built a train simulator and tested it against a railway control system. This simulator achieved good results in simulating the behavior of a moving train using different testing scenarios.

Keywords— *Internet of things; location-based services; vehicle simulator model; train simulator.*

I. INTRODUCTION

Internet of things starts to affect every life aspect and many applications have been implemented in the past few years that change people lifestyle. If we categorize the existing devices into i) connected-devices; where devices has a way to connect and ii) non-connected devices; where devices has no connectivity. IoT enables more devices to be transformed from non-connected into connected world.

The concept is wide-spread to cover not only devices but also solid objects by injecting the required sensors into these objects, for example building a system to monitor the railway bridges to ensure early detection of any bridge damages; this can be achieved by injecting some sensors in the bridge body [1]. The basic concept is the same in all IoT applications; converting things into connected objects that have unique identifiers and are responsible for sharing information or executing an action or both [2].

This enables the efficient utilization of these devices and opens the door for more business applications and better human-machine interactions. If we picked for example the

location-based services, they enable the location tracking of different things that can vary from human beings to vehicles. Therefore we can build different business models around such information e.g. location-based marketing [3].

Other usages such as asset tracking, geo-fencing [4] have become essential parts of car security and traffic monitoring. The concept spread to cover additional areas such as using business intelligent analysis to identify traffic status using the collected data from different tracked vehicles and their own speed in different roads.

To develop business intelligent location-based applications a need to have vehicle simulators that facilitate the development of these applications and testing the business model around the proposed services without the need to invest in the hardware until a complete and mature model become clear, this simulator role is essential before building a new LBS system or altering an existing system to reduce the possibility of failing to fulfill system specifications and also to optimize the system performance [5].

The structure of this paper is as following; in this section we provide background information of the current location-based IoT services, in the next section we will describe the problem definition, after that we will move to discuss the proposed model for building a location-based simulator, we will then discuss the methods and experimental results and we will finish the paper with research conclusion.

II. PROBLEM DEFINITION

The IoT application has become one of the hottest research areas over the past few years and looking at the location-based services many applications has been proposed to solve existing business problems such as asset tracking, geo-fencing, traffic analysis etc. With more involvement of intelligent analysis of collected location data a requirement to have location based simulators that enable the evolution of these services using pure software model has become clear.

This can push the creation of more innovative applications that utilize the data generated by the vehicle simulators. One example is developing a railway control system using the location-based services and other services. To simulate such railway control system a requirement to deploy a hardware device to communicate with this control system in different trains to collect the train's location data. This will not only cost us a lot in terms of devices and connectivity in the early stages of the research but it will also slow down the research progress by the hardware capability constraints. One additional challenge here is the need to upgrade of firmware in these devices with each change in the system in particular in the communication protocol.

All these reasons can point to a clear requirement to develop a location-based simulator so we can speed up the research and reduce the time to market in LBS and therefore reduce the overall solution cost. It also enables the flexibility of developing the communication protocol, hardware and device features, etc. The purpose of this research is to develop a generic customizable location-based simulator model that can be used in developing location-based services and in particular for train location tracking.

III. EXISTING SIMULATION MODELS

The existence of many location services in the internet world such as Google Maps® and the inclusion of these services in the smart phone world expanded the application domain of LBS. Different simulators exist to simulate these services with different level of simulator maturity that varies from a simple location output to a very advanced street-viewer simulator as in Google maps street viewer.

If we focus on train simulators as an example, many simulators have been developed using different concepts, in the early stages of these simulators, the simulator has to contain the collection of track geometry and related speed restrictions [6]. Because train simulators are much more complex awareness of track signaling status is also required, so if we need to build a train simulator we need to consider having a messaging system to send the signal details to the simulator to reflect the received railway signals in its behavior.

Two simulation models are usually used; time and event-based models, in time-based models; the time is divided into spaced intervals where movement is evaluated at each interval. This is near real simulation model and easier for development but it needs much more computational power which can be reduced by increasing the time intervals.

While in event-based model, the movement calculations occur only in pre-defined events (e.g. train leave or arrive to station) which lead to less computation but inconsistent movement updates, this ideally fits in timetable or traffic control applications [6].

Many train simulators exist and can be used to do railway simulation if we pick one example as Train Operation Model (TOM) which contains three different type of simulation; Train Performance Simulators (TPS) which simulate a single train movement, Train Movement Simulator (TMS) which simulates the performance of multi-train network and Electric Network Simulator (ENS) that simulate power flow in the railway system [7]. This TOM does the simulation in respect to railway system as a whole not only the train as abstract, another simulation model called TrainSim which is limited to calculating the train speed profiles using different speed calculation methods [8], some more advanced 3D simulators exist as Open Rail train simulator; which is powerful train simulator but doesn't fit the purpose of this research which is providing simulator for IoT applications [9].

In our research we used the time-based simulation technique with a configurable time slicing that can be tailored according to the application nature and we build a model around this simulation technique that also supports the bidirectional communications (inbound and outbound) to allow complete interaction with any LBS system.

IV. BUILDING VESIMULATOR MODEL

In order to build the simulation model many steps are required; the first step is data collection. Data collection is the most important step as it collects the data that will be fed into the simulator so it uses it to simulate the vehicle movement. The data should be gathered according to the required business model. In our research we have selected train location data as our source to build a train simulator. The following steps describe what we did to develop our location-based simulator:

A. Gathering vehicle location dataset

In this step we have collected the railway location dataset by using a GPS-enabled device while the train is moving from the start station to the destination station. The device logs the location periodically and allows the user to mark certain location such as starting station, crossing areas, middle stations, and final station locations manually. To generate the data for our research we developed an Android® application to get the train locations and log them into a device local file. Most of smart phones have already Assisted-GPS (A-GPS) technology which gives a higher GPS location accuracy by integrating both mobile network and GPS. The add value from A-GPS is that it provides a quick location fix and a better coverage especially inside buildings [10].

In Android OS, there is already support for location APIs that we can use. The main class here is the LocationManager class which is responsible for returning the current device location [11]:

```
LocationManager locationManager = (LocationManager)
getSystemService(Context.LOCATION_SERVICE);
```

After initiating this LocationManager class, we can get the location by calling getLastKnownLocation() method to get the location.

```
Location location = locationManager
.getLastKnownLocation(LocationManager.GPS_PROVIDER);
```

We can also get the location with each change in device location by registering the application main class to listen to these location changes. In the application we have configured the location updates to be with each 100 meters in distance; this can be configured according to the business model and the required location frequency.

```
locationManager.requestLocationUpdates(
    LocationManager.GPS_PROVIDER,
    10000, // 10 seconds
    100, // 100 meters,
    this); // the listener class
```

It is important to define this location change sampling according to our application requirements for example every 50 meters, 100 meters, or 500 meters. In this code we have configured the location update to be fired every 10 seconds or 100 meters difference from the previous location. With each device location change the onLocationChanged(Location) method gets called with the new device location where we need to implement the location logging logic (in a local device file).

```
public void onLocationChanged(Location location) {
    double latitude = location.getLatitude();
    double longitude = location.getLongitude();
    float accuracy = location.getAccuracy();
    // logging logic here ...
}
```

To run this Android application, it will need the permission to access the fine-tuned location of the device as following:

```
<uses-permission
android:name="android.permission.ACCESS_FINE_LOCATION" />
```

Table I contains a sample of collected location data formatted in a table.

TABLE I. SAMPLE OF COLLECTED DATA

Latitude (D)	Longitude (D)	Accuracy	Point type
31.259905	32.300598	30	S
31.259428	32.300255	45	
31.258933	32.299912	30	
31.258492	32.299590	30	
31.257979	32.299246	30	
30.604054	32.300908	30	C
30.603149	32.299878	30	
30.601930	32.298483	45	S

The file as we can see logs the latitude, longitude and accuracy of each location point, the recorded accuracy is very important as it describes the pseudo-range of the reported location, this value depends on many factors such as atmospheric conditions, and GPS device receiver quality, as we will see later this value is important and shouldn't be neglected in most of location-based applications.

The application logs each point with a tag (S, C or nothing) by allowing manual marking of the current location position by either no-mark for normal position and 2 additional mark types either C or S. In our simulator we used C for railway crossing location and S for railway stations. All the logged entries by default are logged without any flag but if the user clicks on station or crossing buttons in the application, the application logs the corresponding flag. In our service we didn't record the altitude of the location as it is not required in our railway system but we can also record it in the collected data if required.

If we are tracing a bus for example the values would be different as we will map the points into bus stations, traffic lights (instead of crossings), squares, etc.

B. Manipulating the location datasets

In this step we manipulate the application output data file to give some meaningful values to the collected data. For instance we can add friendly name to railway stations, update stations flag to distinguish between start station, middle station and final station, add the available number of platforms for each station, and add the max speed allowed for each crossing.

The following three sample lines show the file format after adding such information:

```
31.259905, 32.300598, 30, SS, Railway Station Name, 3
31.241799, 32.298023, 30, C, 40
30.601930, 32.298483, 45, S, Middle Station Name, 2
...
```

The first line describes the station as start station (SS) with 3 platforms, the second line describes the crossing location with maximum speed allowance as 40 Km/Hr, and in the third line we have added middle station name and number of platforms in this station.

Again if we are building simulator for other applications such as bus simulator, the points should always correspond to our simulation milestones with different meanings:

- 31.259905, 32.300598, 30, SS, City Central Bus Station, 8
- 31.241799, 32.298023, 30, C
- 30.601930, 32.298483, 45, S, Town Centre Bus Station, 2

In the first line we have added the start station name and number of bus parking slots as 8. In the third line we have added the station name and number of minutes the bus will wait there (2 minutes).

C. Building the simulator data model

Building the simulation data model is a challenging step to ensure that the system is flexible for future changes especially to add additional relations in the future.

The model is composed from different inter-related tables where the location table represents the core part and gets connected to other tables that represent the purpose and the usage of our simulator. In our railway model the location dataset is linked to railway data, which is represented by a railway database table which in turn get connected to the following tables; stations, crossings, switches, maps, milestones and events. Milestones represents all the location points that are collected in that railway, these points are what the simulator will use to simulate the train movements using vehicle speed to check-in different points.

In fig. 1 we can see part of the simulation data model that we used in our railway application, the location table is in the heart of the application data model and get connected to almost all major tables in this model, we have created a simple application to import the collected data into our system model; during importing the model, the distance between each location milestone is calculated.

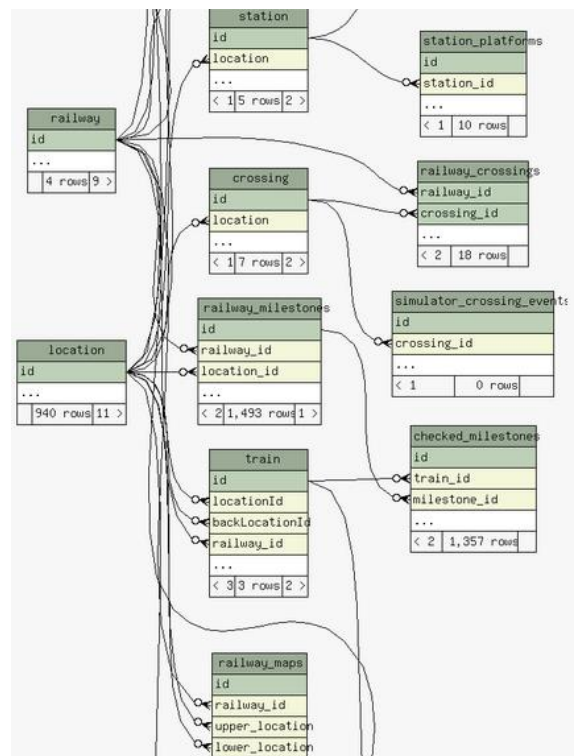


Fig. 1. Part of train simulator data model

The import utility uses the following distance calculation equation to calculate the distance between any two location points A and B using (1).

$$\begin{aligned}
 d = r * & \text{Math.acos}(\text{Math.cos}(\text{Math.toRadians}(\text{latitudeA})) \\
 & * \text{Math.cos}(\text{Math.toRadians}(\text{latitudeB})) \\
 & * \text{Math.cos}(\text{Math.toRadians}(\text{longitudeB}) - \\
 & \quad \text{Math.toRadians}(\text{longitudeA})) \\
 & + \text{Math.sin}(\text{Math.toRadians}(\text{latitudeA})) \\
 & * \text{Math.sin}(\text{Math.toRadians}(\text{latitudeB}))) \quad (1)
 \end{aligned}$$

Where “r” is the distance from the center to the surface of the earth, it is a range not an absolute value because of the asymmetry of the earth sphere, but in our system calculations we fixed its value as 6,357 Km which is the effective radius of the earth at sea-level [12].

In this step we need to supply the simulator with vehicle data as its unique identifier, maximum speed, acceleration, and deceleration as basic information to simulate the vehicle movement. In our train simulator example, the train simulator needs additional information such as railway, current station, train direction, etc.

For other simulation cases we will need other type of information for example in a bus simulator we will need information such as bus identifier, traffic average wait time, max speed allowed for some streets, etc.

D. Adding simulator logic

The simulator initial position is set to the start station position, and then it uses the train speed to calculate the run

distance. Once the train runs the milestone distance, the simulator uses that milestone location as its current location while communicating its location to the LBS system. The simulator needs to implement a communication protocol with the location application. This includes communicating the simulator location together with the simulator unique identifier.

The logic includes implementing different simulator scenarios to cover all possible business scenarios, in train simulator we built different simulation scenarios such as normal scenario, train broken scenario, split car scenario, etc.

The following train simulation scenarios are built using the simulation model by providing different configurations and custom business logic as in Table II:

TABLE II. SOME GENERATED TRAIN SCENARIOS

Scenario	Business Logic
Normal scenario	No custom logic, the simulator will move the train from start station to end station passing through middle stations and reduce speed at crossings.
Train broken scenario	Keep sending same train location at certain point as train speed equals to zero (broken start location).
Train detached cars scenario	Keep sending same train back location after train passes the split location (split start point).
Train exceed max speed scenario	Exceed train max speed after passing certain point.

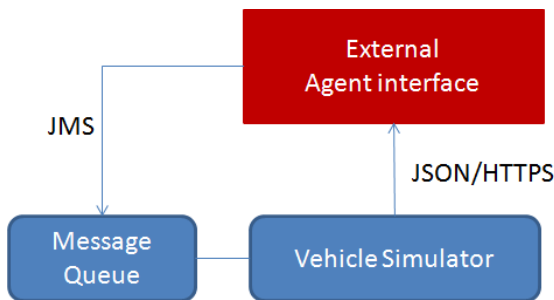


Fig. 2. Simulator inbound/outbound interface

When we execute the simulator we need to feed it up with essential information as initial location, destination location, direction, simulation scenario, speed at each point of time, etc. In our train system the simulator which represents the train will also receives some control agent commands to control the simulator behavior (e.g. speed).

To support this feature we have implemented a messaging queue (using Java Message Service – JMS) where these commands are sent and the simulator keep listening to this messaging queue to execute the received commands, this is important feature to support external control feature of the simulator as shown in fig. 2.

V. VESIMULATOR MODEL

A. The abstract model

The VeSimulator model is composed of the following main components; simulator core, data store (database), controller, interfaces (inbound and outbound) and business logic, fig. 3 shows these different components.

The database contains two main data; the location set and the basic simulator information as id, max speed, initial position, status, acceleration and deceleration. The controller part is responsible for lifecycle management operations such as start, stop, pause and resume of the simulator; it exposes these methods for external systems to control the simulator behavior.

The Inbound/outbound connections are the simulator interfaces with the LBS application, the simulator sends the agreed information in JSON object format to the LBS application and receives the agree action JSON object format. The remaining part in the model is the business logic which is the application specific business logic. The simulator core is responsible for retrieving database data, process the simulator logic and any custom business rules, inbound/outbound connections and applies controller commands. This controller component enables the building of a control dashboard to control the simulator.

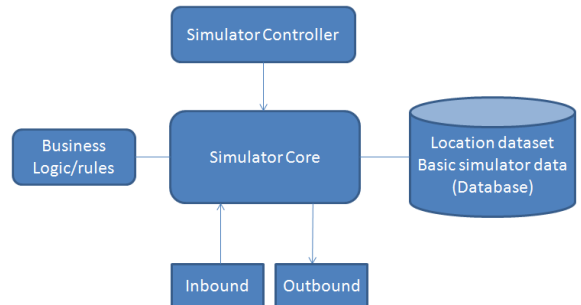


Fig. 3. VeSimulator model

B. Simulator speed calculation

As the simulator simulate the movement speed, some points need to be considered here while calculating the final speed at any moment including acceleration and deceleration power, where both are fed to the simulator according to the average values of our simulator model (e.g. average train acceleration and deceleration).

Also supplying the maximum allowed speed on different situations, so for example we have the following speed limitations; train maximum-allowed speed, crossing speed limit, curvature speed limit, switch maximum speed, external speed limit command, etc. and these different speed limitations may overlap pushing the simulator to pick the minimal allowed speed at any point [8].

As the simulator is configured to run each n seconds, the average speed over that period is used to calculate the vehicle run distance (2) and using this distance the existing location point is determined and sent as the current vehicle location. This means the actual point has a rough margin of error that is

merely dependent on the frequency of location point sampling rate.

$$run_distance = direction * abs(V_f + V_0) * n / 2 \quad (2)$$

Where “V₀“ is the speed at the beginning of n period, “V_f“ is the speed at end of the n period and “direction” is the vehicle direction either forward=1 or backward=-1.

C. Using the simulator

When we deal with location-based services we have to be cautious with the location accuracy. The simulator will keep the accuracy for each location when it communicates it with the application. We should consider the location accuracy and latency in conjugation with current object speed. Having a scenario as we can see in fig. 4; if we have two reported points P1 and P2, both points have an accuracy reported from the GPS device [13]. When we calculate the actual distance between both points we need to consider the worst-case scenario, so for instance if two running vehicles are facing each other the actual distance “d” calculated using (3).

$$d = abs(P1 - P2) - (P1_{accuracy} + P2_{accuracy}) \quad (3)$$

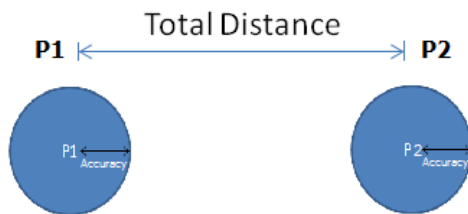


Fig. 4. Total distance between 2 points.

Having IoT location-based applications we need to add one or two additional parameters to our equation to calculate the actual distance; the latency and error buffers. Both values could be configurable and calculated based on the worst-case scenario as well. For example latency buffer which is added to compensate the time spent in device connection to the application, processing time, respond time and execute the action by the device. We can assume this will take 1 second in the average, in that case if the movement speed is 100 Km/Hr the latency is: 100000/3600 = 28 meters. This value represents the run distance by the moving object in that second using its current speed; we need to set this value to a value calculated from the max allowed speed in our business domain to add more safety to the system (e.g. train system).

The error buffer or safety margin is added to the calculations as well to increase the safety of the system or we can combine both in one value. The final equation should look like (4).

$$d = abs(P1 - P2) - (P1_{accuracy} + P2_{accuracy} + latency_buffer + error_buffer) \quad (4)$$

VI. RESULT AND DISCUSSION

We have used VeSimulator to create different train simulators and using these simulators in railway control system, we build different test scenarios to test the application ability to control the train in these different running situations. The control system is validated by testing it with different test cases that are constructed using different combination of these scenarios. The system achieved the required target by simulating the train different scenarios and responding back to the railway control system [14].

According to the application nature these scenarios need to be created with different custom logic inside the simulator model so we can cover many real simulation tests to test our application efficiently.

The VeSimulator supports bi-directional interaction so it sends the location position and receive the response back from the tested system; this enables the implementation of different interactive location-based services.

The simulator was tested as an isolated system using the following aspects; simulator controls (start, stop, pause), simulator custom logic scenarios (scenario execution at specific location) and simulator distance against time; all the simulator test cases passed successfully.

Comparing the VeSimulator features with other existing simulators such as TOM can identify the following aspects as shown in table III.

TABLE III. COMPARISON BETWEEN TOM AND VE-SIMULATOR

Aspect	VeSimulator	TOM
Scope	Vehicle location movement simulator	Whole railway system simulator
Speed calculations	Parameters are externally fed	Parameters are calculated according to the given data
External interface	Available	N/A
Outbound data	Yes	Yes
Inbound data	Yes	No
Configurable	Yes	Yes
Support business rules	Yes	No
Application	IoT LBS	Railway simulation

The simulation model is simple and flexible to be adapted as we did in adapting it for train simulation, it is also configurable system so it can be configured in terms of the selected simulation points distance, simulator accuracy is high as being generated from actual collected data, the simulator uses the time based simulation technique to determine the location and finally data re-calibrations can be done without impacting the built system model by re-importing the data again into the model.

VII. CONCLUSION

The location-based simulator is essential part of developing IoT location-based services; the simulator can improve the

speed in developing different location services especially that require business intelligent analysis. It is a generic simulator that can be simply adapted to any LBS applications.

Using the VeSimulator to develop a railway control system achieved the required bi-directional interaction between the simulator and the control system and enabled us to evolve the system to a mature level with less cost and time.

REFERENCES

- [1] Ying Sun, "Research on the Railroad Bridge Monitoring Platform Based on the Internet of Things" in *International Journal of Control & Automations*, vol.7, no.1, 2014, pp.401-408.
- [2] D. Bandyopadhyay, J. Sen, "Internet of things: Applications and challenges in technology and standardization" in *Wireless Personal Communications*, vol. 58, no. 1, 2011, pp. 49-69.
- [3] J. Schiller and A. Voisard, "General aspects of location-based services" in *Location-Based Services*, Morgan Kaufmann/Elsevier, San Francisco, CA, USA, 2004, ch. 1, pp. 15-32.
- [4] F. Reclus and K. Drouard, "Geofencing for fleet and freight management" in *Intelligent Transport Systems Telecommunications*, ITST, 9th International Conference on, Oct 2009, pp. 353-356.
- [5] A. Maria, "Introduction to modeling and simulation" in WSC '97: Proceedings of the 29th conference on Winter simulation. Washington, DC, USA: IEEE Computer Society, 1997, pp. 7-13.
- [6] GOODMAN C.J., SIU, L.K. and HO, T.K, "A review of simulation models for railway systems" in *International Conference On Development in Mass Transit Systems*, 1998, pp. 80-85.
- [7] Railway System Center, <http://www.railsystemscenter.com/>. last retrieved April, 2015.
- [8] J. C. Jong and S. Chang, "Algorithms for generating train speed profiles" in *Journal of the Eastern Asia Society for Transportation Studies*, vol.6, 2005, pp.356-371.
- [9] Open Rails, <http://www.openrails.org/>. last retrieved April, 2015.
- [10] Manav Singha, Anupam Shukla, "Implementation of location based services in Android using GPS and web services" in *International Journal of Computer Science Issues*, vol. 9, issue 1, no 2, 2012, pp 237-242.
- [11] Location Manager APIs- Android Developer, <http://developer.android.com/reference/android/location/LocationManager.html>. last retrieved April, 2015.
- [12] Witchayangkoon, Boonsap, "Elements of GPS precise point positioning", Diss. University of New Brunswick, 2000.
- [13] Peter H. Dana, "Global positioning system overview", <http://www.colorado.edu/geography/gcraft/notes/gps/gps.html>. last retrieved April, 2015.
- [14] Mostafa Abdel-Azim, Osama Oransa, "Railway as a thing : new railway control system in Egypt using IoT" in *Science and Information Conference*, London, UK, SAI, 2015, in-press.

Numerical Simulation for Fuzzy Fredholm Integral Equations Using Reproducing Kernel Algorithm

Omar Abu Arqub

Department of Mathematics
Faculty of Science
Al Balqa Applied University
Al-Salt, Jordan

E-mail: o.abuarqub@bau.edu.jo

Hasan Rashaideh

Department of Computer Sciences
Faculty of IT
Al Balqa Applied University
Al-Salt, Jordan

E-mail: rashaideh@bau.edu.jo

Shadi Aljawarneh

Software Engineering
Faculty of IT
Al-Isra University
Amman, Jordan

E-mail: shadi.jawarneh@yahoo.com

Abstract- In this paper, we simulate the numerical solutions of fuzzy Fredholm integral equations based on the reproducing kernel algorithm. Using parametric form of fuzzy numbers we convert a linear fuzzy integral equation into a linear system of integral equations in crisp case. The solution methodology is based on generating the orthogonal basis from the obtained kernel functions; whilst the orthonormal basis is constructing in order to formulate and utilize the solutions with series form in terms of their parametric form in an appropriate space. Numerical example is provided to illustrate potentiality of our algorithm for solving such fuzzy equations.

Keywords- Fuzzy Fredholm integral equations; Reproducing kernel algorithm

I. INTRODUCTION

The fuzzy Fredholm integral equations (FFIEs) are important part of the fuzzy analysis theory and they have the important value of theory and application in control theory, measure theory, and radiation transfer in a semi-infinite atmosphere. Generally, many real-world problems are too complex to be defined in precise terms; uncertainty is often involved in any real-world design process. Fuzzy sets provide a widely appreciated tool to introduce uncertain parameters into mathematical applications. In many applications, at least some of the parameters of the model should be represented by fuzzy numbers rather than crisp numbers. Thus, it is immensely important to develop appropriate and applicable algorithm to accomplish the mathematical construction that would appropriately treat FFIEs and solve them.

The aim of this paper is to extend the application of the reproducing kernel Hilbert space (RKHS) method to provide numerical solution for the linear FFIEs of the form

$$x(t) = \int_0^1 h(t, \tau)x(\tau)d\tau, 0 \leq \tau, t \leq 1, \quad (1)$$

where $h(t, \tau)$ is continuous arbitrary crisp kernel functions over the square $0 \leq \tau, t \leq 1$. Here, $\mathbb{R}_{\mathcal{F}}$ denote the set of fuzzy numbers on \mathbb{R} .

Reproducing kernel theory has important application in numerical analysis, computational mathematics, image processing, machine learning, finance, and probability and statistics [1-4]. Recently, a lot of research work has been devoted to the applications of the RKHS method for wide classes of stochastic and deterministic problems involving

operator equations, differential equations, integral equations, and integro-differential equations. The RKHS method was successfully used by many authors to investigate several scientific applications side by side with their theories. The reader is kindly requested to go through [5-13] in order to know more details about the RKHS method, including its history, its modification for use, its scientific applications, its symmetric kernel functions, and its characteristics.

The RKHS method possess several advantages; first, it is of global nature in terms of the solutions obtained as well as its ability to solve other mathematical, physical, and engineering problems; second, it is accurate and need less effort to achieve the results; third, in the RKHS method, it is possible to pick any point in the interval of integration and as well the approximate solution will be applicable; fourth, the method does not require discretization of the variables, and it is not effected by computation round off errors and one is not faced with necessity of large computer memory and time.

Recently, the numerical solvability of FFIEs has been studied by several authors using different numerical or analytical methods. The reader is asked to refer to [14-17] in order to know more details about these analyzes and methods, including their kinds and history, their modifications and conditions for use, their scientific applications, their importance and characteristics, and their relationship including the differences.

The organization of the paper is as follows. In the next section, we present some necessary definitions and preliminary results from the fuzzy calculus theory. The

procedure of solving FFIEs is presented in section III. In section IV, reproducing kernel algorithm is built and introduced. Numerical algorithm and simulation results are presented in Section V. This article ends in section VI with some concluding remarks.

II. FUZZY CALCULUS THEORY

Fuzzy calculus is the study of theory and applications of integrals and derivatives of uncertain functions. This branch of mathematical analysis, extensively investigated in the recent years, has emerged as an effective and powerful tool for the mathematical modeling of several engineering and scientific phenomena. In this section, we present some necessary definitions from fuzzy calculus theory and preliminary results.

Let X be a nonempty set, a fuzzy set u in X is characterized by its membership function $u: X \rightarrow [0,1]$. Thus, $u(s)$ is interpreted as the degree of membership of an element s in the fuzzy set u for each $s \in X$. A fuzzy set u on \mathbb{R} is called convex, if for each $s, t \in \mathbb{R}$ and $\lambda \in [0,1]$, $u(\lambda s + (1 - \lambda)t) \geq \min\{u(s), u(t)\}$, is called upper semicontinuous, if $\{s \in \mathbb{R}: u(s) > r\}$ is closed for each $r \in [0,1]$, and is called normal, if there is $s \in \mathbb{R}$ such that $u(s) = 1$. The support of a fuzzy set u is defined as $\{s \in \mathbb{R}: u(s) > 0\}$.

Definition II.1 [18] A fuzzy number u is a fuzzy subset of the real line with a normal, convex, and upper semicontinuous membership function of bounded support.

For each $r \in (0,1]$, set $[u]^r = \{s \in \mathbb{R}: u(s) \geq r\}$ and $[u]^0 = \overline{\{s \in \mathbb{R}: u(s) > 0\}}$. Then, it easily to establish that u is a fuzzy number if and only if $[u]^r$ is compact convex subset of \mathbb{R} for each $r \in [0,1]$ and $[u]^1 \neq \emptyset$ [19]. Thus, if u is a fuzzy number, then $[u]^r = [u_1(r), u_2(r)]$, where $u_1(r) = \min\{s: s \in [u]^r\}$ and $u_2(r) = \max\{s: s \in [u]^r\}$ for each $r \in [0,1]$. The symbol $[u]^r$ is called the r -cut representation or parametric form of a fuzzy number u .

Theorem II.1 [19] Suppose that the functions $u_1, u_2: [0,1] \rightarrow \mathbb{R}$ satisfy the following conditions; first, u_1 is a bounded increasing and u_2 is a bounded decreasing with $u_1(1) \leq u_2(1)$; second, for each $k \in (0,1]$, u_1 and u_2 are left-hand continuous at $r = k$; third, u_1 and u_2 are right-hand continuous at $r = 0$. Then $u: \mathbb{R} \rightarrow [0,1]$ defined by $u(s) = \sup\{r: u_1(r) \leq s \leq u_2(r)\}$, is a fuzzy number with parameterization $[u_1(r), u_2(r)]$. Furthermore, if $u: \mathbb{R} \rightarrow [0,1]$ is a fuzzy number with parameterization $[u_1(r), u_2(r)]$, then the functions u_1 and u_2 satisfy the aforementioned conditions.

In general, we can represent an arbitrary fuzzy number u by an order pair of functions (u_1, u_2) which satisfy the requirements of Theorem II.1. Frequently, we will write

simply u_{1r} and u_{2r} instead of $u_1(r)$ and $u_2(r)$, respectively.

The metric structure on $\mathbb{R}_{\mathcal{F}}$ is given by $d_{\infty}: \mathbb{R}_{\mathcal{F}} \times \mathbb{R}_{\mathcal{F}} \rightarrow \mathbb{R}^+ \cup \{0\}$ such that $d_{\infty}(u, v) = \sup_{r \in [0,1]} \max\{|u_{1r} - v_{1r}|, |u_{2r} - v_{2r}|\}$ for arbitrary fuzzy numbers u and v . It is shown in [20] that $(\mathbb{R}_{\mathcal{F}}, d_{\infty})$ is a complete metric space.

Definition II.2 [19] Suppose that $x: [0,1] \rightarrow \mathbb{R}_{\mathcal{F}}$, for each partition $\wp = \{t_0^*, t_1^*, \dots, t_n^*\}$ of $[0,1]$ and for arbitrary points $\xi_i \in [t_{i-1}^*, t_i^*]$, $1 \leq i \leq n$, let $\mathfrak{R}_{\wp} = \sum_{i=1}^n x(\xi_i)(t_i^* - t_{i-1}^*)$ and $\Delta = \max_{1 \leq i \leq n} |t_i^* - t_{i-1}^*|$. Then the definite integral of $x(t)$ over $[0,1]$ is defined by $\int_0^1 x(t) dt = \lim_{\Delta \rightarrow 0} \mathfrak{R}_{\wp}$ provided the limit exists in the metric space $(\mathbb{R}_{\mathcal{F}}, d_{\infty})$.

Theorem II.2 [19] Let $x: [0,1] \rightarrow \mathbb{R}_{\mathcal{F}}$ be continuous fuzzy-valued function and put $[x(t)]^r = [x_{1r}(t), x_{2r}(t)]$ for each $r \in [0,1]$. Then $\int_0^1 x(t) dt$ exist, belong to $\mathbb{R}_{\mathcal{F}}$. x_{1r} and x_{2r} are integrable functions on $[0,1]$, and $[\int_0^1 x(t) dt]^r = \int_0^1 [x(t)]^r dt = [\int_0^1 x_{1r}(t) dt, \int_0^1 x_{2r}(t) dt]$.

III. SOLVING FFIEs

In this section, we study FFIEs using the concept of Riemann integrability in which the FFIE is converted into equivalent system of crisp integral equations (CIEs). Furthermore, an efficient computational algorithm is provided to guarantee the procedure.

Next, FFIE (1) is first formulated as an ordinary set of integral equations, after that, a new discretized form of FFIE (1) is presented. Anyhow, in order to apply our RKHS algorithm, we set $H(t, \tau, x(\tau)) = h(t, \tau)x(\tau)$, further, we write the fuzzy function $x(t)$ in term of its r -cut representation forms to get $[x(t)]^r = [x_{1r}(t), x_{2r}(t)]$. By considering the parametric form for both sides of FFIE (1), one can write

$$[x(t)]^r = \int_0^1 [H(t, \tau, x(\tau))]^r d\tau,$$

where $[H]^r = [H_{1r}, H_{2r}]$ in which H_{1r}, H_{2r} are given in the form of $H_{1r} = \min\{h(t, \tau)x_{1r}(\tau), h(t, \tau)x_{2r}(\tau)\}$ and $H_{2r} = \max\{h(t, \tau)x_{1r}(\tau), h(t, \tau)x_{2r}(\tau)\}$.

Prior to applying the RKHS methods for solving FFIE (1) in its parametric form, we suppose that the crisp kernel function $h(t, \tau)$ is nonnegative for $0 \leq \tau \leq c_1$ and nonpositive for $c_1 \leq \tau \leq 1$. Therefore, according to the previous results the FFIE (1) can be translated into the following equivalent form:

$$\begin{aligned} x_{1r}(t) &= \int_0^{c_1} h(t, \tau)x_{1r}(\tau) d\tau + \int_{c_1}^1 h(t, \tau)x_{2r}(\tau) d\tau, \\ x_{2r}(t) &= \int_0^{c_1} h(t, \tau)x_{2r}(\tau) d\tau + \int_{c_1}^1 h(t, \tau)x_{1r}(\tau) d\tau. \end{aligned} \quad (2)$$

Definition III.1 Let $x: [0,1] \rightarrow \mathbb{R}_{\mathcal{F}}$ be continuous fuzzy-valued function. If x satisfy FFIE (1), then we say that x is a fuzzy solution of FFIE (1).

The object of the next algorithm is to implement a procedure to solve FFIE (1) in parametric form in term of its r -cut representation, where the new obtained system consists of two CIEs.

Algorithm III.1 To find the fuzzy solution of FFIE (1), we discuss the following main steps:

Input: The independent interval $[0,1]$, and the unit truth interval $[0,1]$.

Output: The fuzzy solution of FFIE (1) on $[0,1]$.

Step 1: Set $[H]^r = [H_{1r}, H_{2r}]$,

Step 2: Solve the following system of CIEs for $x_{1r}(t)$ and $x_{2r}(t)$:

$$x_{1r}(t) = \int_0^{c_1} h(t, \tau)x_{1r}(\tau)d\tau + \int_{c_1}^1 h(t, \tau)x_{2r}(\tau)d\tau,$$

$$x_{2r}(t) = \int_0^{c_1} h(t, \tau)x_{2r}(\tau)d\tau + \int_{c_1}^1 h(t, \tau)x_{1r}(\tau)d\tau.$$

Step 3: Ensure that the solution $[x_{1r}(t), x_{2r}(t)]$ are valid level sets for each $r \in [0,1]$.

Step 4: Construct the fuzzy solution $x(t)$ such that $[x(t)]^r = [x_{1r}(t), x_{2r}(t)]$ for each $r \in [0,1]$.

Step 5: Stop.

IV. REPRODUCING KERNEL ALGORITHM

In this section, we utilize the reproducing kernel concept in order to construct the reproducing kernel Hilbert space $W_2^m[0,1]$.

Prior to discussing the applicability of the RKHS method on solving FFIEs and their associated numerical algorithms, it is necessary to present an appropriate brief introduction to preliminary topics from the reproducing kernel theory.

Definition IV.1 [4] Let \mathcal{H} be a Hilbert space of function $\phi: \Omega \rightarrow \mathcal{F}$ on a set Ω . A function $K: \Omega \times \Omega \rightarrow \mathbb{C}$ is a reproducing kernel of \mathcal{H} if the following conditions are satisfied. Firstly, $K(\cdot, t) \in \mathcal{H}$ for each $t \in \Omega$. Secondly, $\langle \phi, K(\cdot, t) \rangle = \phi(t)$ for each $\phi \in \mathcal{H}$ and each $t \in \Omega$.

The second condition in Definition IV.1 is called “the reproducing property” which means that, the value of the function ϕ at the point t is reproduced by the inner product of ϕ with $K(\cdot, t)$. Indeed, a Hilbert spaces \mathcal{H} of functions on a nonempty abstract set Ω is called a reproducing kernel Hilbert spaces if there exists a reproducing kernel K of \mathcal{H} .

Definition IV.2 The inner product space $W_2^m[0,1]$ is defined as $W_2^m[0,1] = \{z: z, z', \dots, z^{(m-1)} \text{ are absolutely continuous real-valued function on } [0,1] \text{ and } z^{(m)} \in L^2[0,1]\}$. The inner product and the norm in $W_2^m[0,1]$ are

defined as $\langle z_1(t), z_2(t) \rangle_{W_2^m} = \sum_{i=0}^{m-1} z_1^{(i)}(0)z_2^{(i)}(0) + \int_0^1 z_1^{(m)}(t)z_2^{(m)}(t)dt$ and $\|z_1\|_{W_2^m} = \sqrt{\langle z_1(t), z_1(t) \rangle_{W_2^m}}$, respectively, where $z_1, z_2 \in W_2^m[0,1]$.

The Hilbert space $W_2^m[0,1]$ is called a reproducing kernel if for each fixed $t \in [0,1]$ and any $z(s) \in W_2^m[0,1]$, there exist $K(t, s) \in W_2^m[0,1]$ (simply $K_t(s)$) and $s \in [0,1]$ such that $\langle z(s), K_t(s) \rangle_{W_2^m} = z(t)$.

Theorem IV.1 The Hilbert space $W_2^m[0,1]$ is a complete reproducing kernel and its reproducing kernel function $R_t^m(s)$ can be written as

$$R_t^m(s)|_{s \leq t} = \sum_{i=0}^{m-1} \frac{1}{(i!)^2} t^i s^i + \frac{1}{((m-1)!)^2} \int_0^s (t-\tau)^{m-1} (s-\tau)^{m-1} d\tau$$

$$R_t^m(s)|_{s > t} = \sum_{i=0}^{m-1} \frac{1}{(i!)^2} t^i s^i + \frac{1}{((m-1)!)^2} \int_0^t (t-\tau)^{m-1} (s-\tau)^{m-1} d\tau$$

Definition IV.3 The inner product space $W^m[0,1]$ is defined as $W^m[0,1] = \{z = (z_1, z_2)^T: z_1, z_2 \in W_2^m[0,1]\}$. The inner product and the norm in $W^m[0,1]$ are building as $\langle z(t), w(t) \rangle_{W^m} = \sum_{i=1}^2 \langle z_i(t), w_i(t) \rangle_{W_2^m}$ and $\|z\|_{W^m} = \sqrt{\sum_{i=1}^2 \|z_i\|_{W_2^m}^2}$, respectively, where $z(t), w(t) \in W^m[0,1]$.

To deal with System (2) in more realistic form via the RKHS approach, define the linear operator $v_{ij}: W_2^2[0,1] \rightarrow W_2^1[0,1]$, $i, j = 1, 2$ such that

$$v_{ij}z(t) = \begin{cases} z(t) - \int_0^{c_1} h(t, \tau)z(\tau)d\tau, & i = j, \\ - \int_{c_1}^1 h(t, \tau)z(\tau)d\tau, & i \neq j. \end{cases}$$

Put $0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $X_r = \begin{bmatrix} x_{1r} \\ x_{2r} \end{bmatrix}$, $V = \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{bmatrix}$, and define the mapping $V: W^2[0,1] \rightarrow W^1[0,1]$. Then, System (2) can be written in a new form equivalent to $VX_r(t) = 0$.

V. SIMULATION RESULT

To show behavior, properties, and applicability of the present RKHS method, linear FFIEs will be solved numerically in this section. In the process of computation, all the symbolic and numerical computations are performed by using MAPLE 13 software package.

Algorithm V.1 To approximate the solution $x^n(t)$ of $x(t)$ for FFIE (1), we do the following main steps:

Input: The dependent interval $[0,1]$, the unit truth interval $[0,1]$, the integers n, m , the kernel functions $R_t^1(s), R_t^2(s)$, the linear operator V , and the crisp kernel functions $h(t, \tau)$.

Output: The RKHS solution $X_r^n(t)$ of $X_r(t)$ for System (2) and thus the RKHS solution $x^n(t)$ of $x(t)$ for FFIE (1).

Step 1: Write $X_r = \begin{bmatrix} x_{1r} \\ x_{2r} \end{bmatrix}$ and $X_r^n(t) = \begin{bmatrix} x_{1r}^n \\ x_{2r}^n \end{bmatrix}$.

Step 2: Fixed t in $[0,1]$ and set $s \in [0,1]$;

If $s \leq t$, set $R_t^2(s) = 1 + ts - \frac{1}{6}t^2(t - 3s)$;

Else set $R_t^2(s) = 1 + ts - \frac{1}{6}s^2(s - 3t)$;

For $i = 1, 2, \dots, n, h = 1, 2, \dots, m, j = 1, 2$:

Set $t_i = \frac{i-1}{n-1}$;

Set $r_h = \frac{h-1}{m-1}$;

Set $\psi_{ij}(t) = \begin{bmatrix} [v_{j1}R_t^2(s)]|_{s=t_i} \\ [v_{j2}R_t^2(s)]|_{s=t_i} \end{bmatrix}$;

Output: the orthogonal function system $\psi_{ij}(t)$.

Step 3: For $l = 2, 3, \dots, n, k = 1, 2, \dots, l$:

Set $d_l = \sqrt{\|\psi_l\|_{W^2}^2 - \sum_{p=1}^{l-1} c_{lp}^2}$;

Set $c_{lk} = \langle \psi_l, \bar{\psi}_k \rangle_{W^2}$;

If $k \neq l$, then set $\beta_{lk} = -\frac{1}{d_l} \sum_{p=k}^{l-1} c_{lp} \beta_{pk}$;

Else set $\beta_{ll} = \frac{1}{d_l}$;

Else set $\beta_{11} = \frac{1}{\|\psi_1\|_{W^2}}$;

Output: the orthogonalization coefficients β_{lk} .

Step 4: For $l = 2, 3, \dots, n - 1, k = 1, 2, \dots, l - 1$:

Set $\bar{\mu}_l(t) = \sum_{k=1}^l \beta_{lk} \mu_k(t)$;

Output: the orthonormal function system $\bar{\mu}_l(t)$.

Step 5: Set $X_{r_h}^0(t_1) = 0$;

For $i = 1, 2, \dots, n$:

Set $\alpha_k = \begin{cases} x_{1r} \left(\frac{t_{k+1}}{2} \right), & k \text{ is odd,} \\ x_{2r} \left(\frac{t_k}{2} \right), & k \text{ is even;} \end{cases}$

Set $X_{r_h}^i(t_i) = \sum_{k=1}^i (\sum_{k=1}^i \beta_{ik} \alpha_k) \bar{\mu}_i(t)$;

Output: the RKHS solution $X_r^n(t)$ of $X_r(t)$.

Step 6: Write $[x^n(t)]^r = [x_{1r}^n, x_{2r}^n]$ to get the RKHS solution in which $[x(t)]^r = [x_{1r}, x_{2r}]$.

Step 7: Stop.

Here, we taking $t_i = \frac{i-1}{n-1}$, $i = 1, 2, \dots, n$ and $r_h = \frac{h-1}{m-1}$, $h = 1, 2, \dots, m$ with the reproducing kernel functions $R_t^1(s)$

and $R_t^2(s)$ on $[0,1]$ in which Algorithms III.1 and V.1 are used throughout the computations.

Example V.1 Consider the following FFIE:

$$x(t) = \frac{1}{3} \pi \sin(\pi t) u + \int_0^1 \pi \sin(2\pi\tau) \sin(\pi t) x(\tau) d\tau,$$

where $0 \leq \tau, t \leq 1$. The exact solution is $x(t) = v\pi \sin(\pi t)$. Here, $[u]^r = [-5r^3 - 2r^2 - 7r + 20, 2r^3 + 5r^2 + 7r - 8]$ and $[v]^r = [-r^3 - r + 4, r^2 + r]$.

Anyhow, for approximating the fuzzy solution, we have the following system of CIEs taking into account that the crisp kernel function $h(t, \tau) = \pi \sin(2\pi\tau) \sin(\pi t)$ is nonnegative on $0 \leq \tau \leq \frac{1}{2}$ and nonpositive on $\frac{1}{2} \leq \tau \leq 1$, regardless the effect of the independent variable t on $[0,1]$:

$$\begin{aligned} x_{1r}(t) &= \frac{1}{3} \pi \sin(\pi t) u_{1r} \\ &+ \int_0^{\frac{1}{2}} \pi \sin(2\pi\tau) \sin(\pi t) x_{1r}(\tau) d\tau \\ &+ \int_{\frac{1}{2}}^1 \pi \sin(2\pi\tau) \sin(\pi t) x_{2r}(\tau) d\tau, \\ x_{2r}(t) &= \frac{1}{3} \pi \sin(\pi t) u_{2r} \\ &+ \int_0^{\frac{1}{2}} \pi \sin(2\pi\tau) \sin(\pi t) x_{2r}(\tau) d\tau \\ &+ \int_{\frac{1}{2}}^1 \pi \sin(2\pi\tau) \sin(\pi t) x_{1r}(\tau) d\tau. \end{aligned}$$

The absolute errors of numerically approximating $x_{1r}(t)$ and $x_{2r}(t)$ for the corresponding CIE system have been calculated for various t and r as shown in Tables 1 and 2. Anyhow, it is clear from the tables that, the approximate solutions are in close agreement with the exact solutions.

VI. CONCLUSION

The study of FFIEs forms a suitable setting for the mathematical modeling of real-world problems in which uncertainty or vagueness pervades. The aim of this paper is to propose a numerical method and the corresponding algorithm to solve linear FFIEs. Numerical results show that the presented method is of higher precision and is easy to apply in programming.

Table 1: The absolute errors of approximating $x_{1r}(t)$ using RKHS method.

t_i	$r = 0$	$r = 0.25$	$r = 0.5$	$r = 0.75$	$r = 1$
0.16	2.2167×10^{-8}	2.0686×10^{-8}	2.7409×10^{-8}	2.7281×10^{-8}	2.3836×10^{-8}
0.32	3.6612×10^{-8}	4.9143×10^{-8}	4.7620×10^{-8}	3.2652×10^{-8}	3.6097×10^{-8}
0.64	7.3751×10^{-8}	6.6190×10^{-8}	6.3982×10^{-8}	6.7446×10^{-8}	6.5208×10^{-8}
0.96	8.3262×10^{-8}	7.9022×10^{-8}	8.1784×10^{-8}	7.2881×10^{-8}	8.0809×10^{-8}

Table 2: The absolute errors of approximating $x_{2r}(t)$ using RKHS method.

t_i	$r = 0$	$r = 0.25$	$r = 0.5$	$r = 0.75$	$r = 1$
0.16	2.3445×10^{-8}	2.1964×10^{-8}	2.8124×10^{-8}	2.3027×10^{-8}	2.6190×10^{-8}
0.32	3.7027×10^{-8}	3.7379×10^{-8}	3.5536×10^{-8}	3.3262×10^{-8}	3.4584×10^{-8}
0.64	6.8755×10^{-8}	6.9376×10^{-8}	6.2167×10^{-8}	6.7450×10^{-8}	6.4224×10^{-8}
0.96	7.6651×10^{-8}	8.8084×10^{-8}	7.5751×10^{-8}	8.6631×10^{-8}	8.1507×10^{-8}

REFERENCES

- [1] H.L. Weinert, "Reproducing Kernel Hilbert Spaces: Applications in Statistical Signal Processing", Hutchinson Ross, 1982.
- [2] A. Daniel, "Reproducing Kernel Spaces and Applications", Springer, Basel, Switzerland, 2003.
- [3] A. Berlinet, C.T. Agnan, "Reproducing Kernel Hilbert Space in Probability and Statistics", Kluwer Academic Publishers, Boston, Mass, USA, 2004.
- [4] M. Cui, Y. Lin, "Nonlinear Numerical Analysis in the Reproducing Kernel Space", Nova Science, New York, NY, USA, 2009.
- [5] O. Abu Arqub, "An iterative method for solving fourth-order boundary value problems of mixed type integro-differential equations", Journal of Computational Analysis and Applications, In press.
- [6] O. Abu Arqub, M. Al-Smadi, "Numerical algorithm for solving two-point, second-order periodic boundary value problems for mixed integro-differential equations", Applied Mathematics and Computation 243, 2014, pp. 911-922.
- [7] O. Abu Arqub, M. Al-Smadi, S. Momani, "Application of reproducing kernel method for solving nonlinear Fredholm-Volterra integro-differential equations", Abstract and Applied Analysis, Volume 2012, Article ID 839836, 16 pages, 2012. doi:10.1155/2012/839836.
- [8] O. Abu Arqub, M. Al-Smadi, N. Shawagfeh, "Solving Fredholm integro-differential equations using reproducing kernel Hilbert space method", Applied Mathematics and Computation 219, 2013, pp. 8938-8948.
- [9] O. Abu Arqub, S. Momani, S. Al-Mezel, M. Kutbi, "A novel iterative numerical algorithm for the solutions of systems of fuzzy initial value problems", Applied Mathematics & Information Sciences, In Press.
- [10] M. Al-Smadi, O. Abu Arqub, A. El-Ajuo, "A numerical method for solving systems of first-order periodic boundary value problems", Journal of Applied Mathematics, Volume 2014, Article ID 135465, 10 pages, 2014. doi:10.1155/2014/135465.
- [11] M. Al-Smadi, O. Abu Arqub, S. Momani, "A computational method for two-point boundary value problems of fourth-order mixed integro-differential equations", Mathematical Problems in Engineering, Volume 2013, Article ID 832074, 10 pages, 2012. doi:10.1155/2013/832074.
- [12] S. Momani, O. Abu Arqub, T. Hayat, H. Al-Sulami, "A computational method for solving periodic boundary value problems for integro-differential equations of Fredholm-Volterra type", Applied Mathematics and Computation 240, 2014, pp. 229-239.
- [13] N. Shawagfeh, O. Abu Arqub, S. Momani, "Analytical solution of nonlinear second-order periodic boundary value problem using reproducing kernel method", Journal of Computational Analysis and Applications 16, 2014, pp. 750-762.
- [14] H.S. Goghary, M.S. Goghary, "Two computational methods for solving linear Fredholm fuzzy integral equation of the second kind", Applied Mathematics and Computation 182, 2006, pp. 791-796.
- [15] E. Babolian, H.S. Goghary, S. Abbasbandy, "Numerical solution of linear Fredholm fuzzy integral equations of the second kind by Adomian method", Applied Mathematics and Computation 161, 2005, pp. 733-744.
- [16] S. Abbasbandy, E. Babolian, M. Alavi, "Numerical method for solving linear Fredholm fuzzy integral equations of the second kind", Chaos, Solitons and Fractals 31, 2007, pp. 138-146.
- [17] A. Molabahrami, A. Shidfar, A. Ghyasi, "An analytical method for solving linear Fredholm fuzzy integral equations of the second kind", Computers and Mathematics with Applications 61, 2011, pp. 2754-2761.
- [18] O. Kaleva, Fuzzy differential equations, "Fuzzy Sets and Systems" 24, 1987, pp. 301-317.
- [19] R. Goetschel, W. Voxman, "Elementary fuzzy calculus", Fuzzy Sets and Systems 18, 1986, pp. 31-43.
- [20] M.L. Puri, "Fuzzy random variables", Journal of Mathematical Analysis and Applications 114, 1986, pp. 409-422.

Modeling and Simulation of Electroactive Polymer Robotic Actuator

Md. Masum Billah, Raisuddin Khan and Amir Akramin Shafie

Department of Mechatronics Engineering, Faculty of Engineering
International Islamic University Malaysia
53100 Kuala Lumpur, Malaysia
mdmasum.b@live.iium.edu.my

Abstract— Flexible snake robotic actuator inspired by the performance of snakes together with muscular and vertebrate, well in obstacle constrained rough environment and are capable of complex motions. These types of actuator possess a wide range of motion while also achieving complex geometrical configurations. Although, flexible structures that mimic muscular actuation like the snake links have been attempted in the literature by using shape memory alloys (SMAs), or strings and cables, light-weight, relatively power-dense dielectric electroactive polymers (EAP) can also be used in unison with a flexible snake robot actuator structure to provide actuation. This paper presents a theoretical and experimental study of the Dielectric EAP (DEAP) actuator for designing flexible snake robot actuator. DEAP is silicon artificial muscles that will actuated by applying voltage which perform a great actuation due to its linear large deflection. Preliminary efforts have been made to develop a prototype actuator design as well as learn about the EAP material properties through experimentation.

Keywords—*electroactive polymer; robot; actuator; muscle.*

I. INTRODUCTION

Snake robots are of special interest to both robotics engineers and materials scientists because they combine extreme flexibility (they can bend at any point) with a capability for executing various sophisticated tasks, such as small hole navigation, manipulating small objects etc [1,2]. Engineered flexible structures that mimic muscular robots have been attempted in the literature, an actuator was demonstrated [3,4], by using pneumatic air muscles (PAM), which exhibits disadvantages of less precision, actuation delays, and a bulky compressed air generator. Also, PAM-based actuators cannot be used for light-weight robots. In essence, artificial muscles can rival biological counterparts have yet to be reported.

In an attempt to mimic the functionality of a muscular body by achieving similar high local curvature and complex configurations, different designs for a articulated snake robot actuator design that utilizes dielectric electro-active polymers (EAPs) as the artificial muscles have been explored. Compare with other materials used in the literature, EAPs are the only viable option to build a lightweight actuator for highly compliant artificial muscles. Polymer-based actuator materials are expected to work better than piezoelectric ceramic (PZT) actuators and shape-memory alloys (SMAs) in terms of the amount of strain achieved during actuation (0.1% for PZTs [5]) and the electromechanical coupling that limits the large-scale maneuverability respectively [6]. They can also provide power densities within a factor of 3-5 of the electric motors but with the added flexibility of their muscle-like nature. Furthermore,

dielectric EAPs are good candidates for actuation because they exhibit quick response times and are capable of relatively large strains (10-100%). DEAPs are commercially available in the form of ribbons and sheets, and are typically prestrained on compliant frames (up to 500%) to reduce the actuation voltage. Prestraining is the act of stretching the elastomer, in either a unilateral direction or bilaterally, to reduce the thickness of the material.

Significant work has been done to explore the potential of the EAPs in actuator technologies while attempting to model the behaviour of this material. Out-of-plane actuation has been looked at and implemented using an agonist-antagonist configuration of the EAPs [7], and finite element methods (FEM), as well as analytical modelling of a circular in-plane EAP actuator has been examined [8]. The viscoelastic properties of VHB4910, a commercially available dielectric EAP, have also been characterized in the modelling process [9]. Different actuation methods like the spring roll actuator, and the contractive EAP actuator have been designed, developed, and manufactured to show their effectiveness [9,10,11]. Although there is extensive prior work in developing and modelling actuators separately, few works have been done to consolidate this technical know-how to design a truly flexible and complaint robotic actuator that resembles an flexible snake robot.

In this paper, preliminary steps have been taken in this direction of a design concept for this type of flexible actuator while utilizing a hyper-elastic material model to assist in the design process.

This research is supported by Science grant, Ministry of Science, Technology and Innovation (MOSTI) Malaysia.

II. EAP BACKGROUND

A number of conventional actuator technologies are being used today which include hydraulic, pneumatic, and electromechanical actuators. However, one common drawback that conventional actuators have is that the actuator itself or the power source is bulky and heavy. The need for light-weight, compliant, small-sized actuators yet having high power to weight ratio has driven research into active materials. Dielectric EAP consist of a thin film of elastomer sandwiched between to compliant electrodes and essentially are compliant capacitors that actuate when a DC voltage potential shown in Fig. 1 is applied across the electrodes. The electrodes have opposite charges hence attracted each other and thereby squeezing the thin film of elastomer. Since the elastomer is incompressible, it expands in the planar direction and thins in the thickness direction. Also, the same charges on an electrode repel each other leading to further expansion in the planar direction. This squeezing stress is called a compressive Maxwell stress. The areas that are covered in electrodes are called the active zones. This actuation strain is the source of motion for any actuator made from EAPs. However, most actuators must also utilize the prestrain of the EAP to allow for more efficient use of the motion. Generally, the larger the prestrain is, the lower the voltage required to activate the EAP. The tension-compression type of actuation can be seen from various actuators like a push-pull actuator, and a hinge actuator [7].

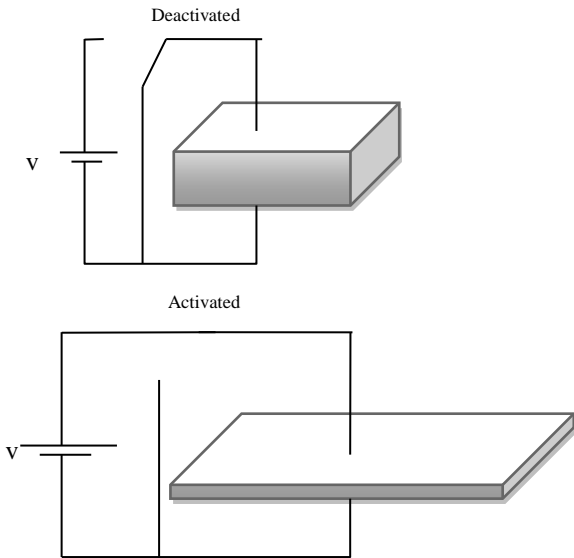


Fig. 1. Dielectric EAP actuator in deactivated (left) and activated state (right).

The previously mentioned actuators all use the extension of the EAPs to provide actuation; however, some actuators are designed to take advantage of the reduction in thickness of the EAP to provide actuation. These actuators require multiple

layers of the elastomer to visibly observe the contraction. The active zones alternate in polarity for each layer so that each layer of the actuator is squeezed. Widely-used dielectric elastomers (DEs) are made out of either silicone or acrylic. VHB 4910 is acrylic EAPs that can strain more and are capable of more tensile stress during Prestraining (increased elastic energy density) when compared to silicone EAPs, however, silicone EAPs have a faster response time and are more efficient mechanically [12].

III. ANALYSIS OF DEAP

A. Constitutive Models

The purpose of the tensile tests were to determine the material parameters of the VHB 4910 DEAP that would be used in the modelling the behaviour of the material and will provide vital information on designing and predicting the movement of specific snake robot link designs. These material parameters are constants in the strain energy function denoted by W . The general form of the strain energy function of a hyper-elastic material subject to a axial stress is denoted by equation (1) [13].

$$W = \sum_{a=0}^3 \sum_{b=0}^3 C_{ab} (I_1 - 3)^a (I_2 - 3)^b \quad (1)$$

$$I_1 = \lambda_1^2 + \lambda_2^2 + \lambda_3^2 \quad (2)$$

$$I_2 = \lambda_1^2 \lambda_2^2 + \lambda_2^2 \lambda_3^2 + \lambda_1^2 \lambda_3^2 \quad (3)$$

I_1 and I_2 denote the first and second invariant of the so-called left Cauchy-Green stress tensor. λ_i ($i = 1; 2; 3$) are the principal strain ratios. Note that I_1 and I_2 are functions of the principal strain ratios. 1; 2; or 3 denotes the direction on the local elastomer sample in which the strains and stresses act. 1 denotes direction in which the tensile stress for the uniaxial test acts, 2 is the direction perpendicular to 1 in the plane of the material, and 3 is the direction resulting from the cross product of 1 and 2 is shown in Fig. 2. C_{ab} are the material parameters in which the uniaxial tests were performed to determine. Based on the strain-energy function, equations (3) and (4) are used for an incompressible, hyperelastic material to determine the principal stresses in terms of the principal strain ratios.

$$\sigma_{11} = \lambda_1^2 \frac{\partial W}{\partial I_1} + \lambda_1^2 (\lambda_2^2 + \lambda_3^2) \frac{\partial W}{\partial I_2} + \text{Phydro} \quad (4)$$

$$\sigma_{22} = \lambda_2^2 \frac{\partial W}{\partial I_1} + \lambda_2^2 (\lambda_3^2 + \lambda_1^2) \frac{\partial W}{\partial I_2} + \text{Phydro} \quad (5)$$

$$\sigma_{33} = \lambda_3^2 \frac{\partial W}{\partial I_1} + \lambda_3^2 (\lambda_1^2 + \lambda_2^2) \frac{\partial W}{\partial I_2} + \text{Phydro} \quad (6)$$

$$\lambda_1 \lambda_2 \lambda_3 = 1 \quad (7)$$

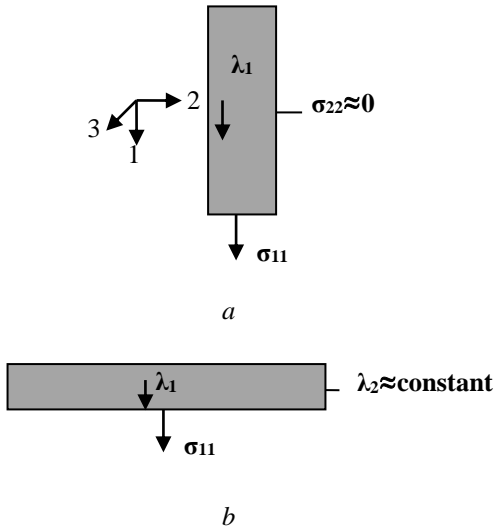


Fig. 2. Elastomer sample (a) Long and thin, (b) Wide-strip

$\sigma_{11}, \sigma_{22}, \sigma_{33}$ denotes the normal stress acting on the material in the 1, 2, and 3 directions and Phydro is the hydrostatic pressure. Equation (7) is the incompressible material condition which can be assumed for the VHB4910 elastomer. For boundary conditions specific to the uniaxial test, the experimental data was fitted to the first equation (4) in the 1 direction. For instance, for a long, thin elastomer strip, the boundary conditions were $\sigma_{22} = 0, \sigma_{33} = 0$. Figure 1(a) shows what a uniaxial test with a long, thin elastomer sample would look like.

Long, thin elastomer samples are samples in which the original, unprestrained length is much greater than the width. This was done to reduce the amount of necking in the sample so that the assumptions for the boundary conditions of the simulation would be a good approximation (i.e. $\sigma_{22} = 0$). Wide-strip elastomer samples are samples in which the original, unprestrained width is much greater than the length. The assumptions for the boundary conditions of this sample are that the prestrain in the 2 direction remains constant ($\lambda_2 = \text{const}$) and $\sigma_{33} = 0$. Based on the correct boundary conditions, the stress-to-strain relationship was derived using equations (7) and (8) and simplified to equations (5).

$$\sigma_{11} = \left(\lambda_1^2 - \frac{1}{\lambda_1^2 \lambda_2^2} \right) \frac{\partial W}{\partial I_1} + \left(-\frac{1}{\lambda_1^2} + \lambda_1^2 \lambda_2^2 \right) \frac{\partial W}{\partial I_2} \quad (8)$$

$$\sigma_{22} = \left(\lambda_2^2 - \frac{1}{\lambda_1^2 \lambda_2^2} \right) \frac{\partial W}{\partial I_1} + \left(-\frac{1}{\lambda_2^2} + \lambda_1^2 \lambda_2^2 \right) \frac{\partial W}{\partial I_2} \quad (9)$$

Equation (5) is used to analyze both the long, thin elastomer and wide-strip elastomer samples, however, the correct boundary conditions need to be applied. For instance, when determining σ_{11} in terms of λ_1 for a long, thin elastomer sample, σ_{22} was set to zero and λ_2 was solved for to get an expression in terms of λ_1 . This was plugged back into the σ_{22} equation to get an expression only in terms of λ_1 . For analyzing a wide-strip sample, the σ_{22} equation was ignored because the σ_{11} equation is already in terms of λ_1 . λ_2 was simply the prestrain of the sample in the 2 direction since it was assumed to be constant. Both types of samples were tested and simulated. Once the experimental data was collected, the data was fitted to a stress-strain model and the material parameters were determined. Then, simulations using these material parameters were done to examine the disparity between the simulation results and the experimental data.

B. Theoretical Analysis of Linear EAP Actuators

In the designs of snake robot actuator, a linear EAP actuator is utilized or the actuation can be approximated by a linear EAP actuator. This means that upon activation of the EAP, the resulting strain that occurs is dominated in one direction (denoted by the 1 direction). Equations (11) show the constitutive model of the linear EAP actuator and were used to draw some insight on the design of an actuator.

$$\sigma_{11} = \lambda_1^2 \frac{\partial W}{\partial I_1} + \lambda_1^2 \left(\lambda_2^2 + \frac{1}{\lambda_1^2 \lambda_2^2} \right) \frac{\partial W}{\partial I_2} + \text{Phydro} = \frac{F_1}{y^1 \lambda_3 z_0} \quad (10)$$

$$\sigma_{22} = \lambda_1^2 \frac{\partial W}{\partial I_1} + \lambda_2^2 \frac{\partial W}{\partial I_1} + \lambda_2^2 \left(\lambda_1^2 + \frac{1}{\lambda_1^2 \lambda_2^2} \right) \frac{\partial W}{\partial I_2} + \text{Phydro} \quad (11)$$

$$\sigma_{33} = \frac{1}{\lambda_1^2 \lambda_2^2} \frac{\partial W}{\partial I_1} + \frac{1}{\lambda_1^2 \lambda_2^2} (\lambda_1^2 + \lambda_2^2) \frac{\partial W}{\partial I_2} + \text{Phydro} = \frac{\epsilon_r \epsilon_0 V^2}{(\lambda_3 z_0)^2} \quad (12)$$

This model includes the compressive stress that effectively "squeezed" the EAP together in the thickness direction which is denoted by the right hand side of the third equation. y^1 is the width of the actuator after prestrain, ϵ_r is the free-space dielectric permittivity ($\epsilon_r = 8.85 \cdot 10.12 \text{ F/m}$), ϵ_0 is the relative permittivity of the dielectric material ($\epsilon_0 = 4.7$ for VHB 4910 [14]), V is the voltage difference supplied to the electrodes, and z_0 is the original thickness of the actuator when it is not activated. For some linear actuator designs, there may be free edges such that the actuator is not attached to a rigid structure at every edge. This causes necking and, as a result, the width of the actuator varies along the actuation direction. Solving for the hydrostatic pressure, the following equation is obtained.

$$\text{Phydro} = \frac{\epsilon_r \epsilon_0 \lambda_1^2 \lambda_2^2 V^2}{z_0^2} - \frac{1}{\lambda_1^2 \lambda_2^2} \frac{\partial W}{\partial I_1} + \frac{1}{\lambda_1^2 \lambda_2^2} (\lambda_1^2 + \lambda_2^2) \frac{\partial W}{\partial I_2} \quad (13)$$

This expression is substituted back into the first equation to yield the following:

$$F_1 = \frac{y^I z_0}{\lambda_1 \lambda_2} \left[\lambda_1^2 \frac{\partial W}{\partial I_1} + \lambda_1^2 (\lambda_2^2 + \frac{1}{\lambda_1^2 \lambda_2^2}) \frac{\partial W}{\partial I_2} - \frac{\epsilon_r \epsilon_0 \lambda_1^2 \lambda_2^2 V^2}{z_0^2} - \frac{1}{\lambda_1^2 \lambda_2^2} \frac{\partial W}{\partial I_1} - \frac{1}{\lambda_1^2 \lambda_2^2} (\lambda_1^2 + \lambda_2^2) \frac{\partial W}{\partial I_2} \right] \quad (14)$$

$$\sigma_{11} = \lambda_1^2 \frac{\partial W}{\partial I_1} + \lambda_1^2 \left(\lambda_2^2 + \frac{1}{\lambda_1^2 \lambda_2^2} \right) \frac{\partial W}{\partial I_2} - \frac{\epsilon_r \epsilon_0 \lambda_1^2 \lambda_2^2 V^2}{z_0^2} - \frac{1}{\lambda_1^2 \lambda_2^2} \frac{\partial W}{\partial I_1} - \frac{1}{\lambda_1^2 \lambda_2^2} (\lambda_1^2 + \lambda_2^2) \frac{\partial W}{\partial I_2} \quad (15)$$

Since the stress invariants can be written in terms of just λ_1 and λ_2 from the incompressible assumption, the force of the linear actuator in the 1 direction (direction of dominate strain) is only a function of y^I , z_0 , λ_1 and λ_2 for a given material and a given activation voltage ($F_1 = f(y^I, z_0, \lambda_1, \lambda_2)$). Equation (14) also tells that the force of the EAP actuator will vary linearly with the width and the thickness of the actuator. Also, the width of the actuator is determined by the prestrain in that direction so the force is also implicitly dependent in the prestrain in this manner. This physically means that the width, thickness, and prestrain of the actuator determines how much force is available for the arm to utilized, but the length of the actuator (dimension of the actuator in the activation direction) has no effect. However, if the linear actuator has free edges, then the length will have an effect on the force of the actuator. For instance, longer actuators for a given width would have a less significant necking effect. The necking of the free edges effectively decrease the prestrain of the actuator in that direction and since the force is a function of prestrain, the force would change with different degrees of necking.

Examining equation (14), the stress in the activation direction is solely a function of prestrain given an activation voltage and thickness. It is not a function of the size of the actuator therefore; theoretically, a large EAP actuator and a small EAP actuator, regardless of the width to length ratio, with equal prestrain and thickness should have the same internal stresses during activation or inactivation. However, for linear actuators with free edges and using the same analysis as previously mentioned, the internal stresses would be different for a longer actuator (for a given width) because of a decreased necking effect. This leads to the conclusion that only linear EAP actuators subject to free edges with the same width-to-length ratio and same prestrain can have equal internal stresses given a specific strain. Furthermore, this leads to the fact that the coefficients of the strain energy function, also known as the material parameters of the material, should be the same for these actuators. As a consequence, determining these material parameters with uniaxial tests will allow for stress and strain prediction of different sized but similar-shaped EAP actuators.

C. Theoretical Data Analysis for Long, Thin Elastomer Sample

First, uniaxial tensile simulation were done, thin strip of elastomer (refer to Fig. 1 (up)). From the gathered data points, the deflection measurements and loads needed to be converted into principal strains and stresses in order for the first equation in (8) to be fitted and the material parameters to be determined. So, for the experimental data analysis, the principal stress in the 1 direction was calculated by equation (9).

$$\sigma_{11} = \frac{P}{A} \quad (16)$$

A denotes the cross sectional area of the sample in the non-necking region, and P denotes the amount of force acting on the sample in the elongation direction calculated by $P = mg$. The cross sectional area was calculated by multiplying the measured width of the non-necking region and the thickness. The thickness was calculated using the incompressible assumption equation (7). The principal strain (λ_1) is equal to $\frac{x}{x_0}$ where x is the instantaneous length of the elastomer sample in the 1 direction and x_0 is the initial length of the elastomer in the elongation direction corresponding to the zero prestrain.

The theoretical data was converted from force to stress and from deflection to strain, Matlab's curve-fitting toolbox was used to determine the material parameters corresponding to the first equation in (8). For one experiment with initial prestrain of $\lambda_1 = 1:1389$, $\lambda_2 = 1$ and an initial length in the 1 direction of 6.75 mm, the material parameters for the strain energy model were determined and are shown in Table 1.

The simulation process is done with the boundary conditions ($\sigma_{22} = \sigma_{33} = 0$) for long, thin elastomer strip samples. The superscript I refers to the pre-strained configuration whereas the II refers to the loaded configuration. y^I is the width (in direction 2). z_0 is the original thickness of the elastomer sample prior to prestrain which is 1 mm. This simulation process was performed for loads ranging from the minimum to maximum load. The nominal strain was calculated using equation (17).

$$s = \left(\frac{\lambda_1^{II}}{\lambda_1^I} - 1 \right) 100\% \quad (17)$$

TABLE I. MATERIAL PARAMETERS FOR A LONG, THIN ELASTOMER SAMPLE

Material Parameter	Fitted Value (Pa)
C01	-9.215
C02	274.9
C03	-1.264
C10	-6.944
C11	-50.02
C12	267.6

C13	0.4922
C20	165.9
C21	0.432
C22	-19.28
C23	-0.01165
C30	300.1
C31	-0.8353
C32	0.2947
C33	6.147e-1

each data point. An average width would then be calculated from that parabolic width distribution and this width would be assumed to be constant along the elongation direction of the elastomer sample. This product of the average width and the corresponding thickness calculated by the incompressible assumption equation (4) was the approximated cross sectional area. Fig. 4 shows this approximation. y_{avg}^I would be less than y^I . This approximation was then made for each data point and the y_{avg}^I for each data point was different because the minimum width was different for each data point. The principal stress was then calculated by equation (14). The principal strains (λ_1) were calculated by $\frac{x}{x_0}$ where x is the current length of the elastomer sample in the elongation direction and x_0 is the initial length of the elastomer corresponding to the zero prestrain. Table 2 shows the material parameters for a wide-strip elastomer samples with initial lengths of 21.7 mm, 25.4 mm, 33.8, and widths of 63.5 mm, 76.2 mm, and 101.6 mm respectively. The samples were subject to an initial prestrain of $\lambda_1 = 3$, $\lambda_2 = 5$ (3 in the actuation direction) with an equal width-to-length ratio of about 3. It was evident from the uniaxial tests that the stress-strain curves for all three samples are similar and warrants the fact that the material parameters are the same for different sized actuators with equal width-to-length ratios and equal prestrains (given the error in measurement). The 3 by 5 prestrain of the sample was chosen as it is the prestrain configuration that has shown the best stress and strain properties in terms of using the elastomer as a linear actuator [8]. The material parameters were determined by fitting experimental data to the first equation in (8).

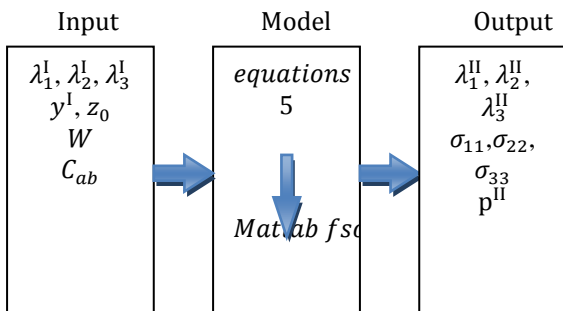


Fig. 3. Simulation process

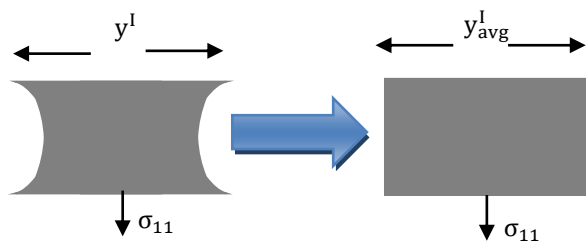


Fig. 4. Pictorial representation of the necking approximation

D. Theoretical Data Analysis for Wide-Strip Elastomer Samples

The wide-strip elastomer sample was analyzed differently from the long, thin elastomer because the boundary conditions are different. The pre-strained length of the wide-strip elastomer sample was much smaller than the pre-strained width. On the other hand, the pre-strained length of the long, thin elastomer was much greater than the pre-strained width. This was the fundamental difference. It was a good approximation in the long, thin elastomer case to assume that the stress in the 2 direction (σ_{22}) was negligible and can be set to zero. This was not a good assumption in this case. However, assuming that the principal strain in the 2 direction (λ_2) was constant turned out to be a much more reasonable assumption. Therefore, only the first equation in equation (8) was needed. To determine the principal stresses and strains from the experimental data, some approximations were made. The necking was approximated geometrically by a parabolic shape determined by the minimum width of the elastomer sample at

TABLE II. MATERIAL PARAMETERS FOR A WIDE STRIP ELASTROMER SAMPLE

Material Parameter	Fitted Value (Pa)
C01	1.2738
C02	2.1262
C03	0.8228
C10	-3.774
C11	1.5459
C12	-4.1087
C13	0.0033
C20	8.1668
C21	-2.7571
C22	-14.3687
C23	-0.0264
C30	0.2192
C31	1.1898
C32	0.6432
C33	1.0505e-5

IV. RESULTS AND DISCUSSIONS

The design of dielectric elastomer is such a way that it can perform large variety of motions depends on the different arrangements of dielectric elastomer strip. We proposed a novel design of elongating elastomer ribbon that will be used as the tendon. The model exhibits linear motion along the axis. The elongating characteristics appear once the dielectric elastomer is electrically activated. The VHB 4910 transparent polymer from 3M, an available EAP for commercial purpose, is found in the market. The ribbon of DE tendons is considered 30 mm long in length. The stress-strain investigation of the material is shown in Fig. 5 (a), (b).

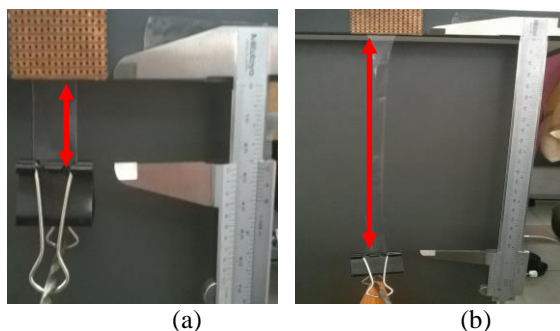


Fig. 5. Elongation test for VHB4910 ribbon: (a) initial length without (b) final length with 380 gm load

The elastomer ribbon's actuation properties are elaborated as stress and strain function of the applied voltage as shown in Fig. 6. Besides the quadratic dependency on the applied voltage and the intrinsic non-linear stress-strain characteristic, the actuator performance is increased by the applied prestrain increases for the development of the displacement range and force.

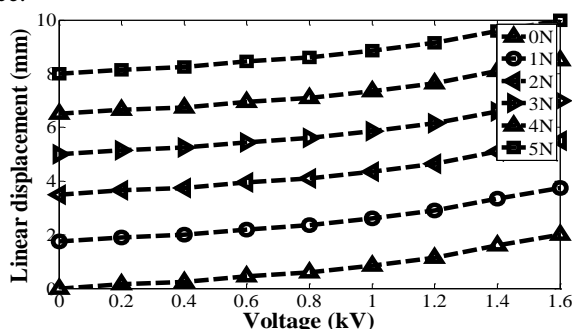


Fig. 6. Voltage-elongation relationships of one of the elastomer actuators.

V. CONCLUSIONS

A variety of designs for flexible snake robot actuator while utilizing an EAP model-guided approach to assist in actuator design is analysed in this paper. Though both long, thin samples and wide-strip samples were characterized, it was later determined that only the wide-strip elastomer configuration was practical for use as an actuator to produce force and stroke.

Therefore, most of the subsequent designs for a flexible arm were based on a wide-strip EAP as its main actuator.

Dielectric elastomer EAPs are promising robotics actuators in bio-inspired artificial robot. They show good mechanical performance, are very low weight, simple and low cost. An important characteristic of DEs is that their polymeric nature makes their performance, reliability, and efficiency highly stretch rate dependant. In particular, the best material known to date for DEAs, VHB 4910, shows improved performance and reliability when operating at high stretch rates. Furthermore, An investigation has been made here are for the tendon driven platform. This step requires further developments, but at the very least, what is evident is the viability, to a limited degree of the proposed design. As the guiding priority is to keep the design small and simple, a successful experimental test would certainly be promising. This development will enhance the inspiration of the researchers to produce more lightweight micro scale actuator in future.

REFERENCES

- [1] W. Kier and M.P. Stella, "The arrangement and function of octopus arm musculature and connective tissue", *Journal of Morphology*, 2007, vol. 268, pp.831-843.
- [2] Y. Yekutieli, G. Sumbre, T. Flash, and B. Hochner, "How to move with no rigid skeleton? The octopus has the answers", *Biologist*, 2002, vol. 49(6), pp. 250-254
- [3] I. D. Walker, D. M. Dawson, T. Flash, F. W. Grasso, R. T. Hanlon, B. Hochner, W. M. Kier, C. C. Pagano, C. D. Rahn, Q. M. Zhang, "Continuum robot arms inspired by cephalopods", *Proc. SPIE*, 2005, Vol. 5804, pp. 303.
- [4] D. Trivedi, C.D. Rahn, W.M. Kier, and I.D. Walker, "Soft robotics: biological inspiration, state of the art, and future research", *Advanced Bionics and Biomechanics*, September 2008, Vol. 5, No. 2, pp. 99-117.
- [5] D. B. Camarillo, C. F. Milne, C. R. Carlson, M. R. Zinn, and J. K. Salisbury, 2008, "Mechanics modeling of tendon-driven continuum manipulators", *IEEE Trans. on Robotics*, 24(6):1262-1273.
- [6] K. J. Cho and H. H. Asada, "Architecture design of a multiaxis cellular actuator array using segmented binary control", *IEEE Transactions on Robotics*, Vol. 22, 2006, no. 4, pp. 831-843.
- [7] P. Lochmatter, "Development of a shell-like electroactive polymer (EAP) actuator", *D-MAVT*. Zuerich, ETH Zurich. Dr. sc. techn. 2007, pp. 17-21.
- [8] M. Wissler, E. Mazza, "Modeling and simulation of dielectric elastomer actuators", *Smart Materials and Structures*, 2005, No. 14, ISSN. 13961402.
- [9] G. Kovacs, L. Doring, "Contractive tension force stack actuator based on soft dielectric EAP", *Proc. of SPIE*, 2009, Vol. 7287, ISSN. 72870A-1.
- [10] F. Carpi, C. Salaris, D. De Rossi, 2007, "Folded dielectric elastomer actuators", *Smart Materials and Structures*, 2007, No. 16, ISSN. S300S305.
- [11] Q. Pei, M. Rosenthal, S. Stanford, H. Prahlah, R. Pelrine, "Multiple-degrees-of- freedom electro elastomer roll actuators", *Smart Materials and Structures*, 2004, No. 13, pp. 86-92.
- [12] P. Brochu, Q. Pei, "Advances in dielectric elastomers for actuators and artificial muscles", *Macromolecular Rapid Communications*, 2010, No. 31, pp. 10-36.
- [13] A. E. Green, W. Zerna, "Theoretical Elasticity", *Oxford University Press*, 2002, ISBN 0-486-49507-8. CH 3.
- [14] G. Kofod, P. Sommer-Larsen, R. Kornbluh, R. Pelrine, "Actuation response of polyacrylate dielectric elastomers", *J. Intell. Mater. Syst. Struct.*, 2003, No. 14, pp. 78-93.

Introduction of Modeling Complex Management Systems using Fuzzy Cognitive Map

Miklos F. Hatwagner

Dept. of Information Technology

Adrienn Buruzs

Dept. of Environmental Engineering

Andras Torma

Dept. of Environmental Engineering

Laszlo T. Koczy

Dept. of Information Technology

Szechenyi Istvan University, Hungary

Abstract— Fuzzy Cognitive Maps (FCM) have found favor in a variety of theoretical and applied contexts that span the hard and soft sciences. Given the utility and flexibility of the method, coupled with the broad appeal of FCM to a variety of scientific disciplines, FCM have been appropriated in many different ways and, depending on the academic discipline in which it has been applied, used to draw a range of conclusions about the belief systems of individuals and groups. In scenario planning, causal mapping has long been used as a means to elicit the worldviews of multiple experts, facilitate discussion, and challenge and improve mental models. Large and complex causal maps, however, are difficult to analyze. The strength of FCM approach lies in its capacity not only to comprehensively model qualitative knowledge which dominates strategic decision making but also to stimulate and evaluate several alternative way of using IT in order to improve organizational performance. This approach introduces computational modeling, as well as it supports scenarios development and simulations. In this paper the authors focus on the investigation of two possible applications: waste management system and stakeholder management system. The common features of these systems are that both systems are complex and comprehensive.

Keywords—Fuzzy Cognitive Maps, Simulation, Integrated Waste Management Systems, Stakeholder Relationship Management

I. INTRODUCTION

Modeling dynamic systems can be hard in a computational sense and many quantitative techniques exist. Well-understood systems may be open to any of the mathematical programming techniques of operations study. First, developing the model usually requires a big deal of effort and specialized knowledge outside the area of interest. Secondly, systems involving important feedback may be nonlinear, in which case a quantitative model may not be possible [1].

This paper presents Fuzzy Cognitive Maps as an approach in modeling the behavior and operation of complex systems. This technique is the fusion of the advances of the fuzzy logic and cognitive maps theories, they are fuzzy weighted directed graphs with feedback that create models that emulate the behavior of complex decision processes using fuzzy causal relations.

Fuzzy Cognitive Maps are fuzzy structures that strongly resemble neural networks, and they have powerful and far-reaching consequences as a mathematical tool for modeling complex systems.

The purpose of this article is to suggest the use of FCM as an alternative approach to existing strategic planning models used in different fields of management. The article suggests that FCM can be a useful tool to facilitate creativity and synergy. There is a wealth of literature from the fields of cognitive science, psychology, and systems science that discusses the use of individuals' knowledge structures as representations or abstractions of real world phenomena [2].

First the description and the methodology that this theory suggests is examined, also some ideas for using this approach in the management area, and then the usage of this tool is described. The application of this approach in the field of system management might contribute to the progress of more intelligent and more objective evaluation of the systems. Fuzzy Cognitive Maps have been fruitfully used in decision making and simulation of complex situation and analysis.

II. FUZZY COGNITIVE MAPS (FCM)

Decision makers usually face serious difficulties when approaching significant, real-world dynamic systems. Such systems are composed of a number of dynamic concepts or

actors which are interrelated in complex ways, usually including feedback links which propagate influences in complicated chains [3].

In the development of the FCM, in the first step of the design process the number and features of constituting factors are determined by the relevant literature. These concepts are supposed to be combined all together in a single system, with mutual interactions.

Modern technological systems are complex and they are usually comprised of a large number of interacting and coupling entities that are called subsystems and/or components. These systems have nonlinear behavior and cannot simply be derived from summation of analyzed individual component behavior [4]. Feedback mechanisms are important in the analysis of vulnerability and resilience of social-economic-technical systems. But how to evaluate systems with direct feedbacks has been a great challenge. FCM was derived from the fusion of fuzzy logic and theory of cognitive maps. Kosko [5] developed the fuzzy signed directed graphs with feedback in order to represent knowledge in a comprehensive way. Since the FCM is formed for a selected system by determining the concepts and their relationships, it is possible to quantitatively simulate the system considering its parameters. It has to be noted however, that a FCM is suitable for short term time series analysis and prediction. A FCM is a dynamic modeling tool in which the resolution of the system representation can be increased by applying a further mapping. The resulting fuzzy model can be used to analyze, simulate, and test the influence of parameters and predict the behavior of the system [6].

According to [6], the design of a FCM is a process that heavily relies on the input from experts and/or stakeholders. This methodology extracts the knowledge from the stakeholders and exploits their experience on the system's model and behavior. A FCM is fairly simple and easy to understand for the participants. With the use of a participatory process it should be ensured that different interests are used to build up synergies as well as partnerships and hence find sustainable solutions as a joint decision [7]. Even though, the cognitive nature of a FCM makes it inevitably a subjective representation of the system. The model is not arbitrary as it is built carefully and reflexively with stakeholders [8].

On the basis of a FCM's development, during the first step in the designing process, the number and features of concepts are determined by a group of experts. After the identification of the main factors affecting the topic under investigation, each stakeholder is asked to describe the existence and type of the

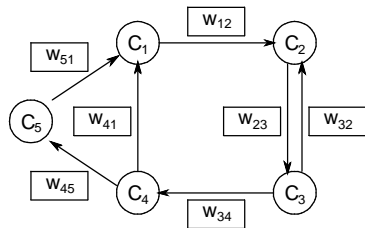


Fig. 1. Example of a simple Fuzzy Cognitive Map

causal relationships among these factors and then assesses the strength of these causal relationships using a predetermined scale, capable to describe any kind of relationship between two factors, positive and negative.

Starting from the primary elements of a FCM, the i th concept denotes a state, a procedure, an event, a variable or an input of the system and is represented by C_i ($i = 1, 2, \dots, n$). Another component of a FCM is the directed edge which connects the concepts i and j . Each edge includes a weight w_{ij} which represents the causality between concepts C_i and C_j . The values of the concepts are within the range $[0, 1]$, while the values of the weights belong to the interval $[-1, 1]$. A positive value of the weight w_{ij} indicates that an increase (decrease) in the value of concept C_i results to an increment (decrement) of the concept's value C_j . Similarly, a negative weight w_{ij} indicates that an increase (decrease) in the value of concept C_i results to a decrement (increment) of the concept's value C_j , while a zero weight denotes the absence of relationship between C_i and C_j (Fig. 1). Considering the interrelations between the concepts of a FCM, the corresponding adjacency matrix can easily be formed. Usually it is accepted that causality is not self reflexive, i.e., a concept cannot cause itself, which means that the weight matrix always has '0-s' in its diagonal [9]. Otherwise the component values may be unstable.

The description of the inference mechanism, which represents the behavior of the physical system, lies in the interpretation of FCM's mathematical formulation. After the initialization of the FCM and the determination of concept activation values by experts, concepts are ready to interact. As it is obvious, the activation of a concept influences the values of concepts that are connected to it. At each step of interaction (simulation step), every concept acquires a new value that is calculated according to equations (Equation 1 and 2) and the interaction between concepts continues until a fix equilibrium is reached; a limit cycle is reached; or a chaotic behavior is observed [10].

The mathematical description of our FCM system is a simple loop:

$$(1)$$

Where V_k is the state k of the system; N is the matrix of the system which contains the weight $W_{i,j}$, and

$$(2)$$

where $\lambda > 0$ determines the steepness of the of the continuous function f .

The FCM is a very convenient and simple tool for modeling complex systems. It is rather popular due to its simplicity and user friendliness. According to [11], human experts are generally rather subjective and can handle only relatively simple networks therefore there is an urgent need to develop methods for automated generation of FCM models.

An FCM is a fuzzy graph structure representing causal reasoning. Causality is represented here as a fuzzy relation of causal concepts. The FCM may be used for dynamic modeling of systems. The FCM approach uses nodes corresponding to the factors and edges for their interactions, to model different aspects in the behavior of the system. These factors interact with each other in the FCM simulation, presenting the dynamics of the original system [4]. The FCM has been described as the combination of neural networks and fuzzy logic. Thus, learning techniques and algorithms can be borrowed and utilized in order to train the FCM and adjust the weights of its interconnections [12].

III. INTEGRATED WASTE MANAGEMENT SYSTEMS

The treatment of waste became one of the most important assignments of today. Several cultural, social, industrial and financial phenomena are responsible for the increasing amount and the more and more diverse types of waste. Many problems of waste processing can be avoided by the consistent usage of source control and appropriate treatment of waste. This way the ratio of reused and recycled waste can be increased. The goal of sustainable waste management is to decrease the amount of waste placed at landfills by e.g. recycling and composting. This part of the paper describes and models the Integrated Waste Management Systems (IWMS) on regional level applying Fuzzy Cognitive Maps (FCM).

During the research process the main factors, the key drivers of a sustainable IWMS were identified at first. After the throughout study of literature [16-21] it can be stated now that a wide-ranging consensus took shape in this topic. The factors are the following: environmental, economic, social, institutional, legal and technical. These factors determine the operation and behavior of such a system. This approach was accepted by the authors and taken into consideration as well-founded.

During the last decades several kinds of models were developed [22] to monitor the processes of waste management, to support decision making and to foresee the possible future outcomes of these decisions. Models based on expert knowledge can help to solve several environmental problems, including IWMS, too. The applied method makes possible to extract the cumulative knowledge and exploit the experiences of stakeholders in order to model the system and its behavior.

The authors' intention was to model an IWMS using FCM. An FCM needs the description of causal relationships among the factors. The factors were already identified using the

TABLE II. THE INITIAL DRAFT OF THE CONNECTION MATRIX

	C1	C2	C3	C4	C5	C6
C1	0	0.8	0.6	0.6	0.4	0.4
C2	0.6	0	0.6	0.6	0.4	0.4
C3	0.8	0.6	0	0.6	0.4	0.4
C4	0.4	0.6	0.4	0	0.4	0.4
C5	0.6	0.6	0.4	0.6	0	0.6
C6	0.4	0.4	0.4	0.4	0.4	0

TABLE I. THE INITIAL STATE VECTOR

Factor	t_0
C1	0.20
C2	0.15
C3	0.10
C4	0.10
C5	0.10
C6	0.10

literature, as it was mentioned before, but the strengths and directions of interactions were unknown. In order to solve this problem, a questionnaire was created. All stakeholders participated in this on-line survey was asked to describe the relationships (directions and weights) between factors using a predetermined simple scale. The applicable values could be both positive and negative. A guideline was also created to support the work of the stakeholders, i.e. to describe the terms of concepts and the goal and basics of research. Finally, 75 different connection matrices were created on the basis of the stakeholders' answers. They were merged into a single but representative map (see Table 1) by averaging source matrices. The factors were denoted in the following way: C1) technical factor, C2) environmental factor, C3) economic factor, C4) social factor, C5) legal factor and C6) institutional factor.

The description of causal relationships among factors is not enough to begin the simulation of IWMS using FCM, however. The other input of the model is the initial state vector of the factors. The data originates from literature [14-21] and represented by real numbers between 0 and 1 (see Table 2).

Several simulations were made with different λ (threshold function parameter) values, but it affects only the values of factors at the end of simulation, not the order of them. This can be an important issue in practice, however, because an unfavorably selected λ results in almost equal factor values that makes really hard to determine the real order of factors. The presented simulation contained 10 iterative steps, but in most cases less iteration would be enough. Fig. 2 shows that factor values converged really fast to their final, stable values. The values of factors in the last simulation step and their order are presented in Table 3. The members of the set containing the

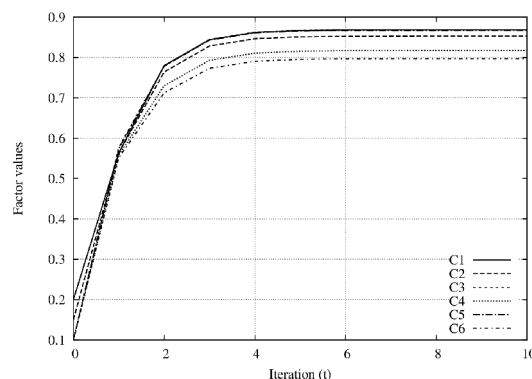


Fig. 2. The model simulation with $\lambda = 0.8$

TABLE III. THE ORDER AND FINAL VALUES OF VECTORS

Factor	Order	Value
C1	2	0.8685
C2	4	0.8530
C3	1	0.8688
C4	5	0.8173
C5	3	0.8675
C6	6	0.7970

highest factor values are C1 (technical factor), C3 (economic factor) and C5 (legal factor). Their values are almost the same. They are followed by factor C2 (environmental factor), C4 (social factor) and C6 (institutional factor) is the last in order.

An important result of the simulation is that the order of factors defines the priority of factors in the IWMS on regional level. The technical factor defines how and what materials are managed, treated and disposed of. This field covers the attributes of collection, transfer and treatment systems, e.g. organic material treatment, thermal treatment, materials recovery and final disposal.

The economic issues (available funding, system costs and revenues, etc.) have practically the same importance. The third component is the relevant legislation, e.g. prescriptive or enabling legislation; EU, national, and municipal level legislation; legal definition of municipal solid waste. The next factor in the order of importance is the environmental factor. It covers e.g. the livability of the settlements, the pollution in different areas. The following element in the list is the social factor. The main issue here is to minimize the risks to public health, adapting the system to the local demands and requirements and to willingness and ability to pay. The final element of the list is the institutional factor. It includes the fields of accountability, stakeholder involvement, transparency and professionalism.

It must be emphasized again that the validity of the results and their applicability in practice depends on the input data. Because the data is collected from a wide range of well-known experts, we are convinced that it is usable to plan or establish new IWMSs at least in a more or less closed geographical area, even if it always could contain subjective convictions.

IV. STAKEHOLDER RELATIONSHIP MANAGEMENT

Authors investigated the applicability of FCM method to analyse the interconnections between the main factors of Stakeholder Management System. The investigation showed practical usability of the method for the desired purpose.

Stakeholders influence the operation and the decisions of an organization, but they play also a determinative role during a project, a program or even during various activities. The influence made by stakeholders can be very different. All of the stakeholders are actors who are linked with the organization. These actors can be identified by their interest regarding the operation and decisions. Beside that also the attitude of the

stakeholders is determinative for this group of actors. Stakeholders can influence the success of the activity in diverse ways, the influence made by them can be in definite cases absolute decisively. Stakeholder management has the aim to deal with stakeholder issues and so it can contribute to the successful operation or to an effective completion of a project. Grouping and identifying the stakeholders by definite parameters is the baseline for all further measures [23]. Well-founded management of stakeholders contributes to the success of the company and long-term sustainability of the organization [24].

Every organization has interested parties also from intern and from extern. Different categories of stakeholders can be identified such as e.g. professional associations, other companies, shareholders, authorities, employees, or even the customers.

The connection of stakeholders with the organization is two sided: on the one hand they have great information demand about the operation, on the other hand they are influencing the operation. It is not extraordinary that they are communicating and cooperating with each other. Other effect of the stakeholder's activity is that they can form the circumstances of the operation.

Establishing an effective stakeholder management system needs deeper knowledge about the interested parties. This knowledge helps the decision makers to set up the main elements of a management approach, which can contribute to the more effective operation. The most important topic is to get a detailed picture. For that there is a need of deeper assessing the stakeholder structure of an organization. Identifying the main actors is very important but the priority ranking of them plays also a determinative role. The need of effectivity necessitates the priority ranking of the stakeholders because

organizations have limited resources to deal with that issues and that's why it is essential to know the most significant ones.

The commonly used methods of stakeholder analysis used by companies are mainly inquiry techniques or checklist-surveys. With them decision makers can get a more accurate level of information about the stakeholders or even about the organizational attitude and activities regarding management processes in the organization.

These techniques are mainly suitable to outline the structure of stakeholders in a static way. With the help of these methods the experts are not able to get information about the dynamics within the system and the interconnections between the main drivers.

The authors investigated in former research the opportunity of the dynamic modeling of a mapped stakeholder system [25]. The aim of these research was to develop a stable methodology for that. Analyzing the interconnections between the driver elements and the causality as well as the weights of them can help in better understanding of a SRMS.

The research stated that the Fuzzy Cognitive Map methodology (FCM) is suitable for this modelling purpose. For

the analysis the authors used expert based data on identifying factors of the SRMS.

Traditionally stakeholder analysis is made in two steps. The purposes of this phase of the analysis are the following: (a) to identify the stakeholders; (b) to analyze the relationship with them [26].

The first purpose can be reached by using different management techniques (e.g.: brainstorming of the experts). For the second purpose there are different methods available. Different matrix approaches are used frequently for this analysis. These matrix approaches traditionally investigate two different dimensions, namely the degree of involvement of the stakeholder (from low to high) or their type of influence (opposing or supporting).

This two-dimensional approach of the interconnections between the interested parties gives sufficient information for setting up suitable management strategies. The picture resulting from such static analysis is shows the attitudes of the actors in a definite time moment. The nature of these connections, the casual correspondences of the system and its cross cutting connections are however hidden.

Researches investigating stakeholder management in respect of mainly on the ideal strategies set up on the basis of the two dimensional approach, the concrete activities and applicable management techniques in connection with them.

Stakeholders are differentiated usually by their attitudes. Fig. 3. shows a classic representation of stakeholder groups with different attitudes. Letters from A to J are representing hypothetic stakeholders of an organization. The main strategy in connection with the stakeholders is to manage the group with the greatest influence and with the greatest interest. It means that the focus area of the management activities is the upper right section of Fig. 3.

Current research does not face with the analysis of causal relations between the stakeholders and the characteristics of this system is not in the focus.

The SRMS gives the frames and the main strategic ways of the management actions mentioned before. Effective solution of the problems regarding stakeholder management can be defined also with the help of the SRMS. Other benefit of such an approach, that this gives the possibility of standardized solutions. Knowing the interactive connections between these main driving elements and the dynamics of such connections gives a detailed picture about the whole system.

TABLE IV. IDENTIFIED CAUSAL CONNECTIONS BETWEEN MAIN CONCEPTS (EXAMPLE OF HUNGARY)

	Weights	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
C1	0.9	0	2	0	6	3	2	9	10	10	8
C2	0.5	10	0	2	9	8	8	2	7	8	4
C3	0.6	5	8	0	9	7	6	6	3	4	6
C4	0.9	6	0	8	0	9	10	4	2	3	3
C5	0.6	7	4	3	-1	0	8	9	8	9	8
C6	0.8	4	1	3	4	2	0	3	9	10	7
C7	0.7	0	8	-2	3	4	4	0	3	3	10
C8	0.9	3	6	0	4	2	2	0	0	9	1
C9	1.0	0	0	0	1	0	1	0	6	0	0
C10	0.4	6	7	1	5	1	3	10	3	4	0

The authors investigated the applicability of FCM to model the interconnections between the main criteria of the SRMS and to make conclusions for the application of effective methods to manage the regarding issues.

As the first step the main driver elements of a SRMS were defined. These elements have the biggest influence on the operation efficiency of a SRMS and so they are subject of the management investigations. For this step results of scientific examinations and notions of practitioner business managers was used. The 10 main identified categories were the following: C1: importance of the stakeholder management; C2: allocation of resources; C3: involvement of employees; C4: organization culture; C5: internal regulations; C6: organizational strategy, policy; C7: internal expectations; C8: external expectations; C9: external regulatory instruments; C10: activity of internal stakeholder parties. The categories were also broken down into 48 subcategories to get in the future more appropriate information.

The possible causality and the weights of the connections were measured by interviewing company and scientific experts in Hungary and in Lithuania. The results of the investigation were used as one input data for the FCM-model.

The first detailed analysis was made by using the Hungarian results, where representatives of the private sector and researchers of this field were asked (more than 15 representatives of different business sectors – e.g.: machine manufacturing, service sector and governmental sector).

TABLE VI. VALUES OF FACTORS WITH OPTIMAL LAMBDA ($\lambda = 0.664$)

Factor	Value
C1	0.941
C2	0.967
C3	0.962
C4	0.938
C5	0.958
C6	0.915
C7	0.877
C8	0.817
C9	0.610
C10	0.913

Fig. 3. Stable approach of the stakeholder analysis

The results were averaged and can be seen in Table 4 (values are varying between $[-10; 10]$).

For the first model the authors used a fix connection. The initial weighting of the factors were also obtained by expert questioning (see in the second column of Table 4). This information describes the initial state vector of the factors.

Table 5 shows the results of the analysis, namely the priority order of the concepts (with $\lambda = 1$; from the most determinative one to the less). The most determinative driver of the SRMS is the availability of the resources (C2) for stakeholder issues. The next two factors both are internal drivers, namely: involvement of employees (C3) and the state of internal regulations within the organization (C5). As fourth and fifth the importance of this topic (C1) and the organizational culture (C4) are playing important role. The strategy and policy of the organization (C6) influences only the sixth important role. The last four drivers are the activity of internal parties (C10), the internal expectations (C7) regarding SRMS, in connection with it the external expectations (C8) and the regulatory instruments of external parties (C9). This ranking gives sufficient information to propose new management strategies for stakeholder issues.

Simulations were made by using different λ parameter

TABLE V. PRIORITY LIST OF THE CONCEPTS ON THE BASIS OF MODEL RESULTS (HUNGARIAN VALUES; WITH $\lambda=1$)

Ranking Hungarian values	
C2	0.9956
C3	0.9944
C5	0.9937
C1	0.9893
C4	0.9870
C6	0.9794
C10	0.9777
C7	0.9585
C8	0.9162
C9	0.6784

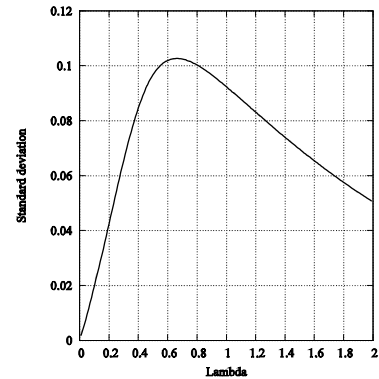


Fig. 4. Value of standard deviation with different λ values

(threshold function parameter). The results showed that the data were inconclusive showing almost the same factor value for different drivers. These values made the practical evaluation practically impossible. Several empirical attempts were made in [27] to find an appropriate value for λ , but the more thorough definition and analysis of this parameter was lacking until now.

Authors looked for the maximal spread out value of them because factor values can be then easier differentiated. The spread will be quantified in the rule of the standard deviation function borrowed from statistics. Different λ values result in different factor values. The standard deviation of factors calculated with different λ values is depicted by Fig. 4. The maximum of the standard deviation is 0.103 at $\lambda = 0.664$. This result was calculated numerically with the Golden Section Search algorithm, a well-known and rather simple kind of line search methods. Using the optimal λ value authors remodeled the causal interconnections. The stable values are listed in Table 6.

Authors used a novel approach to calculate stable FCM values for analyzing management problems. The presented method makes possible to achieve the most easily interpretable simulation results at the cost of executing a computationally inexpensive local search algorithm.

V. FURTHER RESEARCH

Our intention is to validate the developed models by experts of the fields. The expected results of these investigations may help to determine the essential steps towards solving these complex problems on the long term and obtain techniques for the sustainability and long term maintenance of the systems.

ACKNOWLEDGMENT

The authors would like to thank to TÁMOP-4.2.2.A-11/1/KONV-2012-0012, TÁMOP-4.1.1.C-12/1/KONV-2012-0017, to the Hungarian Scientific Research Fund (OTKA) K105529 and K108405 for the support of the research.

REFERENCES

- [1] C. Buche, P. Chevaillier, A. Nédélec, M. Parenthoën, and J. Tisseau. 2010. Fuzzy cognitive maps for the simulation of individual adaptive behaviors. *Comput. Animat. Virtual Worlds* 21, 6 (November 2010), 573-587. DOI=10.1002/cav.363 <http://dx.doi.org/10.1002/cav.363>
- [2] E. I. Papageorgiou (ed.), *Fuzzy Cognitive Maps for Applied Sciences and Engineering*, 29, Intelligent Systems Reference Library 54, DOI: 10.1007/978-3-642-39739-4_2, © Springer-Verlag Berlin Heidelberg 2014
- [3] J. P. Carvalho, J. A. B. Tomé, Rule Based Fuzzy Cognitive Maps - Qualitative Systems Dynamics IFSA-EUSFLAT 2009 Lisbon, Portugal
- [4] Stylos, D. and Groumpos, P. P. (2004). Modelling Complex Systems Using Fuzzy Cognitive Maps. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, 34(1): 155-162.
- [5] Kosko, B. (1986). Fuzzy Cognitive Maps. *Int. J. of Man-Machine Studies*, 24(1): 65-75.
- [6] Papageorgiou, E. and Kontogianni, A. (2012). Using Fuzzy Cognitive Mapping in Environmental Decision Making and Management: A Methodological Primer and an Application. *Int. Perspectives on Global Environmental Change*, S. Young (ed.), ISBN: 978-953-307-815-1, InTech, doi: 10.5772/29375.
- [7] Malena, C. (2004). Strategic Partnership: Challenges and Best Practices in the Management and Governance of Multi-Stakeholder Partnerships Involving UN and Civil Society Actors. Background paper prepared by for the Multi-Stakeholder Workshop on Partnerships and UN-Civil Society Relations, Pocantico, New York.
- [8] Isak, K. G. Q., Wildenberg, M., Adamescu, M., Skov, F., De Blust, G. and Varjopuro, R. (2009). A Long-Term Biodiversity, Ecosystem and Awareness Research Network Manual for Applying Fuzzy Cognitive Mapping – Experiences from ALTER-Net. Project no. GOCE-CT-2003-505298, ALTER-Net Deliverable type: Report, WPR6-2009-02 - Deliverable 4.R6.D2.
- [9] Carvalho, J. P. (2010). On the Semantics and the Use of Fuzzy Cognitive Maps in Social Sciences. *Soft Computing in the Humanities and Social Science*, vol. 214, pp. 6-19.
- [10] Ketipi, M. K., Koulouriotis, D. E., Karakasis, E. G., Papakostas, G. A. and Tourassis, V. D. (2010). A Flexible Nonlinear Approach to Represent Cause-effect Relationships in FCMs. *J. of Applied Soft Computing* 12(12): 3757-3770.
- [11] Stach, W., Kurgan, L., Pedrycz, W. and Reformat, M. (2005). Genetic Learning of Fuzzy Cognitive Maps. *J. of Fuzzy Sets and Systems*, 153(3): 371-401.
- [12] Stylos, C. D., Georgopoulos, V. C. and Groumpos, P. P. (1997). The Use of Fuzzy Cognitive Maps in Modelling Systems. In *Proceedings of 5th IEEE Mediterranean Conf. on Control and Systems*, Paphos, Cyprus.
- [13] A. Demirbas, "Waste Management, Waste Resource Facilities and Waste Conversion Processes", *Energy Conservation and Management* 52, 1280-1287, (2011).
- [14] M. L. M. Graymore, N. G. Sipe, R. E. Rickson, "Regional Sustainability: How Useful are Current Tools of Sustainability Assessment at the Regional Scale?", *Ecological Economics*, Volume 67, Issue 3, 362-372, (2008).
- [15] A. van de Klundert, J. Anschutz, "Integrated Sustainable Waste Management: the Selection of Appropriate Technologies and the Design of Sustainable Systems is Not (Only) a Technological issue", CEDARE/ETC Inter-regional Workshop on Technologies for Sustainable Waste Management, Alexandria, Egypt, 1-17, (1999).
- [16] A. J. Morrissey, J. Browne, "Waste Management Models and Their Application to Sustainable Waste Management", *Waste Management* 24, 297-308, (2004).
- [17] E. J. Wilson, F. R. McDougall, J. Willmore, "Euro-Trash: Searching Europe for a More Sustainable Approach to Waste management", *Resources Conservation and Recycling* 31, 327-346, (2001).
- [18] D. J. Langa, C. R. Binder et al., "Material and Money Flows as a Means for Industry Analysis of Recycling Schemes. A Case Study of Regional Bio-Waste Management", *Resources, Conservation and Recycling* 49, 159-190, (2006).
- [19] J. den Boer, E. den Boer, J. Jager, "LCA-IWM: A Decision Support Tool for Sustainability Assessment of Waste Management Systems", *Waste Management* 27, 1032-1045, (2007).
- [20] S. A. Thorneloe, K. Weitz, M. Barlaz, R. K. Ham, "Tools for Determining Sustainable Waste Management Through Application of Life-Cycle Assessment: Update on U.S. Research", Seventh International Waste Management and Landfill Symposium V, 629-636, (1999).
- [21] S. A. Thorneloe, K. Weitz, M. Barlaz, R. K. Ham, "Tools for Determining Sustainable Waste Management Through Application of Life-Cycle Assessment: Update on U.S. Research", Seventh International Waste Management and Landfill Symposium V, 629-636, (1999).
- [22] E. Papageorgiou, A. Kontogianni, "Using Fuzzy Cognitive Mapping in Environmental Decision Making and Management: A Methodological Primer and an Application", *International Perspectives on Global Environmental Change*, S. Young (ed.), ISBN: 978-953-307-815-1, InTech, DOI: 10.5772/29375, (2012).
- [23] "A guide to the project management body of knowledge (PMBOK® Guide) – Fifth Edition", Project Management Institute, Pennsylvania USA, 2013, pp. 391-413
- [24] F. Perrini, A. Tencati, "Sustainability and Stakeholder Management: the Need for New Corporate Performance Evaluation and Reporting Systems", *Bus. Strat. Env.* 15, 296 – 308, 2006
- [25] M. F. Hatwagner, A. Torma, L. T. Kóczy, Parameter dependence of Fuzzy Cognitive Maps' behaviour, submitted to ASCC 2015, unpublished.
- [26] Team FME, "Project Stakeholder Management – project skills", 2014
- [27] Torma, A., Susniene, D., Hatwagner, F. M., Kóczy, T. L., "A comparative analysis of Stakeholder Management by using FCM", submitted to ICCMIT 2015, unpublished.

UML Activity Diagrams and Maude Integrated Modeling and Analysis Approach Using Graph Transformation

Elhillali Kerkouche, Khaled Khalfaoui
Dept. Computer Science, University of Jijel,
MISC Laboratory, University Constantine 2,
Algeria
{elhillalik, Kh-khalfaoui}@yahoo.fr

Allaoua Chaoui
Dept. Computer Science and its Applications,
MISC Laboratory,
University Constantine 2,
Algeria
chaoui@misc-umc.org

Ali Aldahoud
Al-Zaytoonah University of Jordan,
P.O. Box 130, Amman 11733,
Jordan
aldahoud@zuj.edu.jo

Abstract—The use of UML Activity Diagrams for modeling global dynamic behaviors of systems is very widespread. UML diagrams support developers by means of visual conceptual illustrations. However, the lack of firm semantics for the UML modeling notations makes the detection of behavioral inconsistencies difficult in the initial phases of development. The use of formal methods makes such error detection possible but the learning cost is high. Integrating UML with formal notation is a promising approach that makes UML more precise and allows rigorous analysis. In this paper, we present an approach that integrates UML Activity Diagrams with Rewriting Logic language Maude in order to benefit from the strengths of both approaches. The result is an automated approach and a tool environment that transforms global dynamic behaviors of systems expressed using UML models into their equivalent Maude specifications for analysis purposes. The approach is based on Graph Transformation and the Meta-Modeling tool AToM³ is used. The approach is illustrated through an example.

Keywords— *UML Activity Diagrams; Rewriting Logic; Maude language; Meta-Modeling; Graph Grammars; Graph Transformation; AToM3.*

I. INTRODUCTION

The Unified Modeling Language (UML) [1] has become a widely accepted standard in the software development industry. Some diagrams are used to model the structure of a system while others are used to model the behavior of a system. UML Statecharts, UML collaboration diagrams, UML Sequence Diagrams and UML Activity diagrams are used to model the dynamic behavior in UML. UML State chart diagrams model the lifetime (states life cycle) of an object in response to events. A UML Collaboration diagram models the interaction between a set of objects through the messages (or events) that may be dispatched among them. UML Sequence Diagrams describe an interaction by focusing on the sequence of messages (or events) that are exchanged, along with their corresponding

occurrence specifications on the lifelines. UML Activity diagram model the global dynamic behavior of systems in term of control flow or object flow with emphasis on the sequence and conditions of the flow. UML Activity diagrams are widely used to model workflow systems, service oriented systems and business processes. Control flow includes support for sequential, choice, parallel and events. Activities may be grouped in sub-activities and can be nested at different levels. However, the UML is a semiformal language which lacks rigorously defined constructs.

Rewriting logic has sound and complete semantics [2] and it is considered as one of very powerful logics in description and verification of concurrent systems. Also, the rewriting logic language Maude [3] is considered as one of very

powerful languages based on Rewriting logic. However, Maude system offers textual way to the user to create and deal with systems. Execution under Maude system is done by using command prompt style. In this case, the user loses the graphical notations which are important for the clarity, simplicity and readability of a system description.

In this context, UML and Maude language have complementary features: UML can be used for modeling while Maude can be used for verification and analysis. Thus, developing a tool support for modeling and analysis of complex concurrent systems is significant to modelers who use UML to model their systems. UML behavioral models are projected automatically into Maude specifications for analysis and verification to detect behavioral inconsistencies like deadlock, imperfect termination, etc. Then, the results of the formal analysis can be back-annotated to the UML models to hide the mathematics from modelers.

In this paper, we propose a modeling tool and Graph Transformation approach for modeling and verification of global dynamic behavior in UML models using Maude language. Building a modeling tool from the scratch is a prohibitive task. Meta-Modeling approach is useful to deal with this problem, as it allows the modeling of the formalisms themselves [4]. A model of formalism should contain enough information to permit the automatic generation of tool to check and build models subject to the described formalism's syntax. In order to get a more general transformation approach between UML and Maude, we research the transformation at the Meta-Model level. And for reaching an automatic and correct process, we use Graph Transformation Grammars and Systems to define and implement the transformation. Using our approach, the modelers specify the global dynamics of a system by means of UML Activity diagrams. Then, the modelers transform automatically their behavioral specification into its equivalent Maude specification. From the obtained formal specification, they can use Maude Model Checker to verify their models.

With this end, we have defined a simplified Meta-Model for UML Activity diagrams using ATOM³ tool [5]. Then, we have used this Meta-Modeling tool to automatically generate a visual modeling tool for UML Activity diagrams according to its proposed meta-model. For the transformation process, we have defined a graph grammar to translate the UML Activity diagrams created in the generated tool to a Maude specification. Then the rewriting logic language Maude is used to perform the verification of the resulted Maude specification.

The rest of this paper is organized as follows. Section 2 outlines the major related work. In section 3, we review the main concepts of UML Activity diagrams, Rewriting logic, Maude language and graph transformation. In section 4, we describe our approach that transforms a UML Activity diagrams to Maude specification. In section 5, we illustrate our approach using an example. The final section concludes the paper and gives some perspectives.

II. RELATED WORKS

In the literature, several research works has been done about the integration of different UML diagrams and formal methods such as Petri nets [6] [7] [8], Colored Petri nets (CPN) [9], Object-Z [10], B method [11], LOTOS, Communicating Sequential Processes (CSP) [12] and Maude [13].

For the formalization of UML Activity Diagrams, the most important approaches use CSP or CPN formalisms. In [14], the authors present a case study of UML Activity Diagram to CSP transformation using graph transformation. In [15], the authors describe how an UML activity diagram can be transformed into a corresponding CSP expression by using the graph rewriting language PROGRES. In [16], the author explains how activity semantics are translated into colored Petri net semantics.

On the other hand, the rewriting logic language Maude offers the advantage of its sound and complete semantics [2] and it is considered as one of very powerful languages in specification, programming and verification of non-deterministic concurrent systems. In this paper, UML Activity Diagram semantics are defined in terms of rewriting logic. Rewriting logic gives to UML Activity Diagram a simple, more intuitive and practical textual version to analyze, without losing formal semantic (mathematical rigor, formal reasoning).

III. BACKGROUND

A. UML Activity Diagrams

UML Activity Diagram is one of the important UML models. It is utilized to describe an operation step by step in a system. Moreover, it models the overall control flow between activities and its relationships among several objects with a lot of parallel process. It supports the following concepts: choice, iteration and concurrency. Its structure is a connected graph in which the nodes are represented by icons and the edges by connections. An Activity Diagram includes the following constructs: Initial Node, Flow Final node, Activity Final node, Decision Node, Merge Node, Fork Node Join Node and transition. Only the last construct is represented by a connection; the others are represented by icons. These constructs are shown in Figure 1.

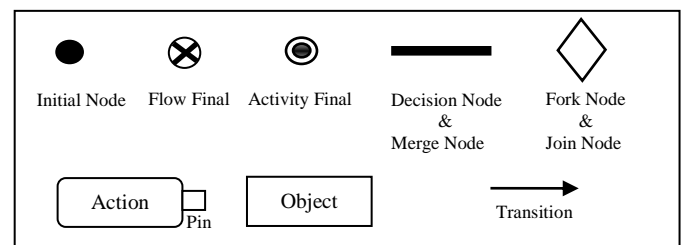


Fig. 1. UML Activity Diagram constructs.

B. Rewriting Logic & Maude Language

In Rewriting Logic, each concurrent system can be specified by a rewriting theory. A rewrite theory is defined as a 4-tuple (Σ, E, L, R) , where the signature (Σ, E) is an equational theory, L is a set of labels and R is a set of possibly conditional labeled rewrite rules that are applied modulo the equations E .

An important consequence of the RL definition is that the rewrite theory can be viewed as an executable specification of the concurrent system that it formalizes. The state is represented by an algebraic term, the transition becomes a rewriting rule and the distributed structure is expressed as an algebraic structure. For more information on the subject see [17].

Maude is a specification and programming language based on Rewriting Logic [18]. It integrates an equational style of functional programming with Rewriting Logic computation. Maude's implementation has been designed with the explicit goals of supporting executable specification and formal methods applications. Three types of modules are defined in Maude specification: The functional modules, the system modules and the object oriented modules. In this work, we will use only functional and system modules

Functional Modules: Functional modules define data types and operations on them by means of equational theories. In other words, Functional modules can be seen as an equational-style functional program with user definable syntax in which a number of sorts, their elements, and functions on those sorts are defined.

System Module: The system module defines the dynamic behavior of a system. It augments the functional modules by the introduction of rewriting rules. A rewriting rule specifies a local concurrent transition which can proceed in a system. The execution of such transition, specified by the rule, can take place when the left part of a rule matches to a portion of the global state of the system and the condition of the rule is valid. This type of module augments the functional modules by the introduction of rewriting rules.

In addition, Maude integrates a model checker. Model-checking is an automatic method for deciding if a specification satisfies a set of properties (for more details, see [19]).

C. AToM³ & Graph Grammar

AToM³ [5] is a visual tool for Multi-formalism Modeling and Meta-Modeling. By means of Meta-Modeling, we can describe or model the different kinds of formalisms needed in the specification and design of systems. Based on these descriptions, AToM³ can automatically generate tools to manipulate (create and edit) models in the formalisms of interest [20].

AToM³'s capabilities are not restricted to these manipulations. AToM³ also supports graph rewriting system, which uses Graph Grammar to visually guide the procedure of model transformation. Graph Grammar [21] is a generalization of Chomsky grammar for graphs. It is a formalism in which the transformation of graph structures can be modeled and studied. The main idea of graph transformation is the rule-based modification of graphs as shown in Fig.1.

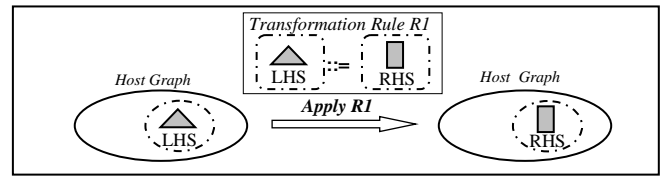


Fig. 2. Rule-based Modification of Graphs.

Graph Grammars are composed of production rules, each having graphs in their left and right hand sides (LHS and RHS). Rules are compared with an input graph called host graph. If a matching is found between the LHS of a rule and a subgraph in the host graph, then the rule can be applied and the matching subgraph of the host graph is replaced by the RHS of the rule. Furthermore, rules may also have a condition that must be satisfied in order for the rule to be applied, as well as actions to be performed when the rule is executed. A graph rewriting system iteratively applies matching rules in the grammar to the host graph, until no more rules are applicable.

IV. OUR APPROACH

The proposed approach consists of transforming a UML Activity diagram to Maude specification. To reach this objective, we have proposed a meta-model for UML activity diagram and a graph grammar that performs automatically the transformation of a UML Activity diagram. In this work, we focus on control flow which addresses the control part of UML Activity diagram, and we leave the object flow for future works. In the following, we describe in details our approach.

A. Meta-modeling

To Meta-model Activity diagrams, we proposed the simplified meta-model containing thirteen classes linked by seven associations and twelve inheritances as shown in Figure 3. Each association of this meta-model links an instance of the source class with a single instance of the destination Class. Some classes are described as follows:

ActionNode Class: represents the Action constructs of the diagram. Graphically it is represented by a rectangle with rounded corners. An Action node has *Name* attribute, and it can be connected with all control nodes, others Action nodes, Object nodes or Pin nodes.

InitialNode Class: represents the beginning of an activity diagram. Graphically it is represented by a small solid circle. It has a constraint which prohibits the existence of incoming Arcs.

To fully define our meta-models, we have also specified the graphical appearance of each entity of the formalisms according to its appropriate graphical notation (shown in Figure 1). Given our meta-model, we have used AToM³ to generate a palate of buttons allowing the user to create the constructs defined in meta-model (see Figure 5).

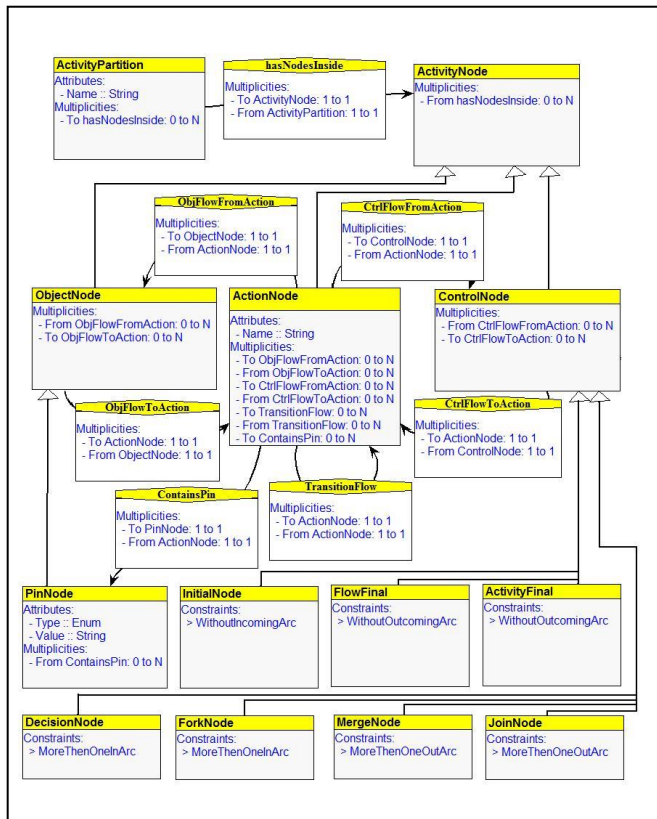


Fig. 3. Simplified UML Activity Diagram Meta-Model

B. Representation of UML Activity Diagram in Maude

In this section, we will explain how to express a UML Activity Diagram in Maude language by using two Modules. We define first a *Basic_ActivityDiagram* functional module that describes basic operations of *Activity Diagram*. This module is described as follows:

```
fmod Basic_ActivityDiagram is

sort CONFIGURATION .
sorts InitialNode ActivityFinal FlowFinal Action .
subsorts InitialNode ActivityFinal FlowFinal Action < CONFIGURATION .
op null : -> CONFIGURATION .
op _ _ : CONFIGURATION CONFIGURATION -> CONFIGURATION [assoc comm id:null] .
op Isin : ActivityFinal CONFIGURATION -> Bool .
vars E E' : ActivityFinal .
vars S conf : CONFIGURATION .
eq Isin (E, Null) = false .
eq Isin (E, E'S) = E==E' or Isin (E, S) .

endfm
```

It contains the declaration of new type called *CONFIGURATION* which represents the current configuration of an Activity diagram instance. The configuration of an Activity diagram consists of Initial Node, Activity Final, Flow Final and/or Actions which are declared as subsorts of *CONFIGURATION*. In addition, this module defines operations used for manipulating configuration elements, as well as equations implementing these operations. For example, The *Isin* operation returns a Boolean value which indicates if Activity Final sub-configuration is in a configuration.

TABLE I. REPRESENTATION OF CONTROL STRUCTURES IN MAUDE

Activity Diagram Control Structures	Corresponding Maude Rewriting Rules
	rl [Initial]: Initial => Act1
	rl [Transition]: Act1 => Act2
	rl [FinalFlow]: Act1 => FinalFlow
	rl [FinalAction]: Act1 => FinalActivity
	rl [Merge]: Act1 => Act4 rl [Merge]: Act2 => Act4 rl [Merge]: Act3 => Act4
	rl [Joint]: Act1 Act2 Act3 => Act4
	rl [DecisionC1]: Act1 => Act2 rl [DecisionC2]: Act1 => Act3 rl [DecisionC3]: Act1 => Act4
	rl [Fork]: Act1 => Act2 Act3 Act4

The second module is *ActivityDiagram* system module that describes transitions firing and control nodes with their conditions (if any) by rewriting rules as shown in Table I.

We note that all rewriting rules (except Initial rewriting rule) are enabled when the Activity Final is not in the current configuration of Activity diagram.

C. Automatic Translation (Graph Grammar)

To generate automatically Maude specification from a UML Activity diagram, we have proposed a Graph Grammar (GG) to traverse the Activity diagram and generate the corresponding code in Maude. The advantage of using a graph grammar to generate the textual code is the graphical and high-level fashion.

The graph grammar has an *initial Action* which opens the file where the code will be generated and decorates all the elements in the model with temporary attributes to be used in the conditions specified in the GG rules. For each element, we use two attributes: *Current* and *Visited*. The *Current* attribute is used to identify the element in the model whose code has to be generated, whereas the *Visited* attribute is used to indicate

whether code for the element has been generated yet. In our GG, we have proposed sixteen rules which will be applied in ascending order by the rewriting system until no more rules are applicable. We are concerned here by code generation, so none of these GG rules will change the Activity diagram models. For lack of space, we only describe the following rules (see Figure 4):

Rule1: Gen_Rule_InitialNode2Action (priority 1): is applied to locate the initial node which is related to an Action node, and generate the corresponding Maude specification.

Rule5: Gen_LeftPartOfForkNodeRule (priority 3): is applied to locate a Fork node which is related to current Action node with an incoming transition, and generate the left part of the corresponding rewriting rule in Maude.

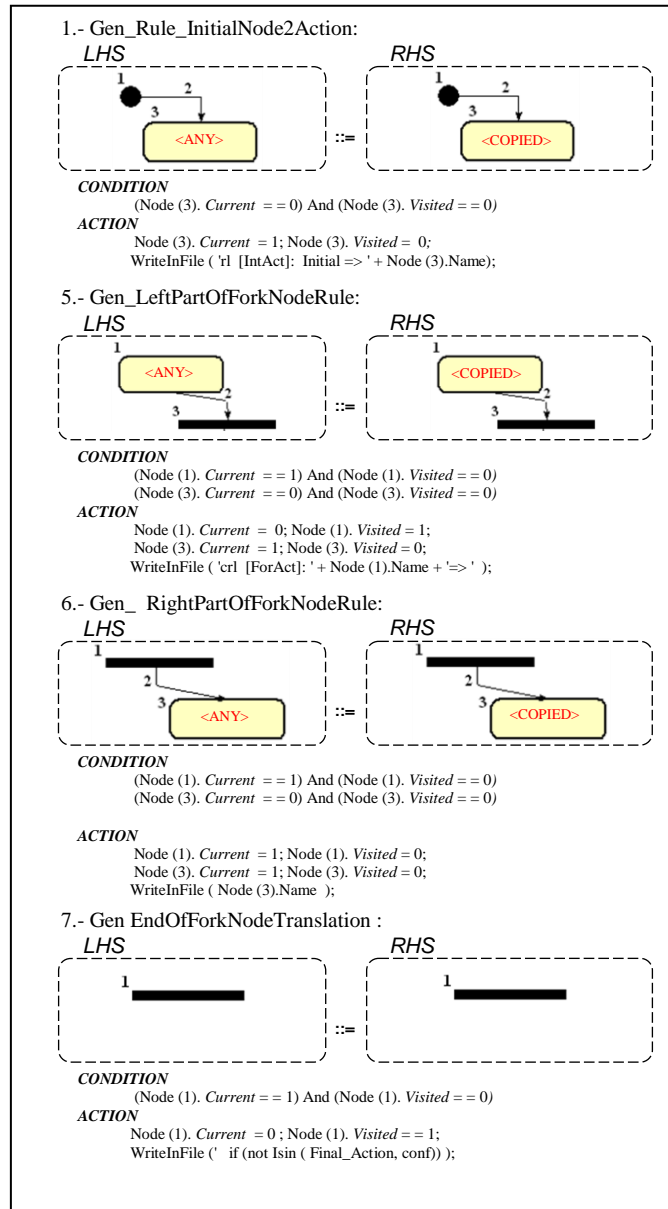


Fig. 4. Graph Grammar to generate Maude specification from an Activity Diagram

Rule6: Gen_RightPartOfForkNodeRule (priority 3): is applied to locate an Action node related to the current Fork node with an incoming transition, and generate its name in the right part of the corresponding rewriting rule in Maude.

Rule7: EndOfForkNodeTranslation (priority 6): is applied to locate the current Fork node whose processing has been terminated, and mark it as Visited. In addition, it generates the condition of the rewriting rule.

The graph grammar has also a final action which erases the temporary attributes from the entities and closes the output file.

V. CASE STUDY

To evaluate the practical usefulness of the proposed approach, we consider a simple example of order processing application. In this diagram, the first action is to receive requested order. After order is accepted and all required information is filled in, payment is accepted and order is shipped. We Note that this example allowing order shipment before invoice is sent or payment is confirmed. The Figure 5 presents the UML Activity diagram of the Process Order created in our tool.

To analyze this behavioral specification of the order processing application, we have to transform this specification into its equivalent Maude specification. To realize this transformation in our tool, we have to execute the proposed Graph Grammar. The resulted Maude specification of the automatic transformation is shown in Figure 6.

In order to perform the analysis by simulation of the resulted Maude specification, we have invoked the rewriting logic Maude system. Simulation consists of transforming the initial state to another by doing one or many rewriting actions. Therefore, in addition to generated file, the user may give to the Simulator the number of rewriting steps if (he/she) wants to check intermediary states. If this number is not given, the Simulator continues the simulation operation until reaching a final state. The Result configuration (final state) of the simulation is given in the same manner as configuration. In our example (see Figure 7), we have asked the application to perform the simulation from the initial node.

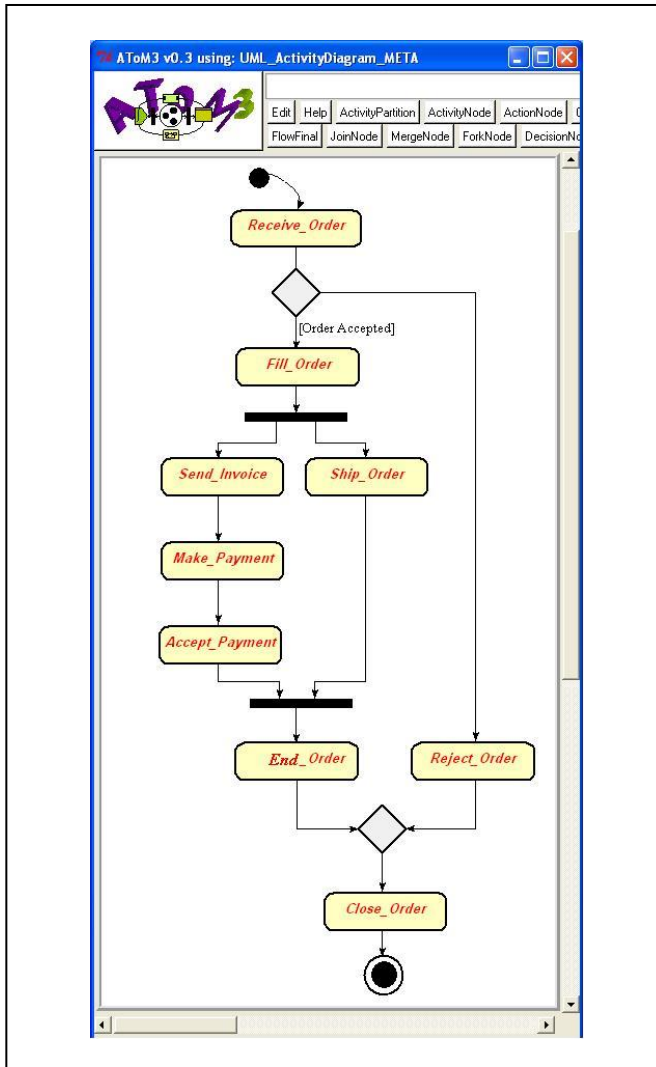


Fig. 5. UML Activity diagram created in our tool

```

ActivityDiagramInMaudeExample: Bloc-notes
Fichier Edition Format Affichage ?
in Basic_ActivityDiagram
mod ActivityDiagram is
op Initial : -> InitialNode .
ops Receive_Order Fill_Order Send_Invoice Ship_Order Make_Payment
    Accept_Payment Close_Order End_Order Reject_Order :-> Action [ctor] .
op FinalAct-Close_Order : -> ActivityFinal .
r1 [Initial]: Initial => Receive_Order .
cr1 [Decision_Accepted]: Receive_Order => Fill_Order if (not isin(FinalAct-Close_Order, conf)) .
cr1 [Decision_Else]: Receive_Order => Reject_Order if (not isin(FinalAct-Close_Order, conf)) .
cr1 [Fork]: Fill_Order => Send_Invoice Ship_Order if (not isin(FinalAct-Close_Order, conf)) .
cr1 [Transition]: Send_Invoice => Make_Payment if (not isin(FinalAct-Close_Order, conf)) .
cr1 [Transition]: Make_Payment => Accept_Payment if (not isin(FinalAct-Close_Order, conf)) .
cr1 [Join]: Ship_Order Accept_Payment => End_Order if (not isin(FinalAct-Close_Order, conf)) .
cr1 [Merge]: Reject_Order => Close_Order if (not isin(FinalAct-Close_Order, conf)) .
cr1 [Merge]: End_Order => Close_Order if (not isin(FinalAct-Close_Order, conf)) .
cr1 [FinalAction]: Close_Order => FinalAct-Close_Order if (not isin(FinalAct-Close_Order, conf)) .
endm
set trace on .
rew Initial .
    
```

Fig. 6. Generated Maude specification

```

Core Maude 2.3
fmod Basic_ActivityDiagram
Done reading in file: "Basic_ActivityDiagram.maude"
mod ActivityDiagram
rewrite in ActivityDiagram : Initial .
rewrites: 41 in 1628036047000ms cpu <0ms real> <0 rewrites/second>
result ActivityFinal: FinalAct-Close_Order
Maude> rewrite in ActivityDiagram : Initial .
rewrite in ActivityDiagram : Initial .
rewrites: 41 in 7383486348ms cpu <0ms real> <0 rewrites/second>
result ActivityFinal: FinalAct-Close_Order
Maude>
    
```

Fig. 7. Execution of order processing example under Maude system.

VI. CONCLUSION

In this paper, we have presented a formal framework and an environment tools based on the combined use of Meta-Modeling and Graph Grammars for the Modeling and analysis of global dynamic behavior in UML models using Maude language. With Meta-modeling, we have defined the syntactic aspect of UML Activity Diagrams, and then we have used the meta-modeling tool AToM³ to generate its visual modeling environment. By means of Graph Grammar, we have extended the capabilities of our framework to transform UML Activity Diagrams into an equivalent Maude specification. The resulted specification can be used to verify system properties using Maude model checking.

In a future work, we plan to transform composite action nodes and complexes links in Maude specification. We plan also to perform some verification of properties using Maude model checking.

REFERENCES

- [1] G. Booch, I. Rumbaugh and J.Jacobson, "The Unified Modeling Language User Guide", in Addison-Wesley, 1999.
- [2] J. Meseguer, "Rewriting Logic as a Semantic Framework of Concurrency: a Progress Report", in Springer-Verlag, Lecture Notes in Computer Science, 119, 1996, pp. 331-372.
- [3] J. Meseguer, "Rewriting logic and Maude: a Wide-Spectrum Semantic Framework for Object-based Distributed Systems", In S. Smith and C.L. Talcott, editors, Formal Methods for Open Object-based Distributed Systems, (FMOODS'2000), 2000, pp. 89-117.
- [4] J. De Lara and H. Vangheluwe, "Meta-Modelling and Graph Grammars for Multi-Paradigm Modelling in AToM³", in Software and Systems Modelling, Special Section on Graph Transformations and Visual Modeling Techniques, Vol. 3, 2004, pp. 194-209.
- [5] AToM³ Home page, <http://atom3.cs.mcgill.ca/>
- [6] J.A. Saldhana, M. Shatz and Z. Hu, "Formalisation of Object Behavior and Interaction From UML Models", in International Journal of Software Engineering and Knowledge Engineering. Vol. 11, #6, 2001, pp. 643-673.
- [7] H. Xinhong, C. Lining, M. Weigang, G. Jinli and X. Guo, "Automatic transformation from UML statechart to Petri nets for safety analysis and verification", Quality, Reliability, Risk, Maintenance, and Safety Engineering (ICQR2MSE), in International Conference on, Conference Publications, Print ISBN: 978-1-4577-1229-6, 2011, pp. 948 - 951.
- [8] M. Wang; L. Lu, "A transformation method from UML statechart to Petri nets", in Computer Science and Automation Engineering (CSAE), 2012 IEEE International Conference on, On page(s): 89 - 92 Vol.2, May 2012, pp.25-27.

- [9] E. Kerkouche, A. Chaoui, E. Bourennane, O. Labbani, "A UML and Colored Petri Nets Integrated Modeling and Analysis Approach using Graph Transformation", In *Journal of Object Technology*, vol. 9, no. 4, 2010, pp 25–43.
- [10] J. Araujo and A. Moreira. "Specifying the Behavior of UML Collaborations Using Object-Z". in *Departamento de Infomatica, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Portugal*, 2000.
- [11] H. Ledang and J. Souquière, "Formalizing UML Behavioral Diagrams with B. Tenth OOPSLA Workshop on Behavioral Semantics: Back to Basics", in Tampa Bay, Florida, USA, 2001.
- [12] C.A.R. Hoare, "Communicating Sequential Processes". In *Prentice Hall International Series in Computer Science*. Prentice Hall, April 1985.
- [13] P. Gagnon, F. Mokhati, M. Badri: "Applying Model Checking to Concurrent UML Models", in *Journal of Object Technology*, Vol. 7, no. 1, January- February 2008, pp. 59-84, http://www.jot.fm/issues/issue_2008_01/article1/
- [14] D. Bisztray, K. Ehrig, and Reiko Heckel, "Case Study: UML to CSP Transformation". Available at <http://www.informatik.uni-marburg.de/~swt/agtivecontest/UML-to-CSP.pdf>
- [15] E. Weinell and U. Ranger, "Using PROGRES for Transforming UML Activity Diagrams into CSP Expressions". Available at www.se.rwthachen.de/files/agtivetc/UML_to_CSP.pdf.
- [16] H. Störrle, "Structured Nodes in UML 2.0 Activities", in *Nordic Journal of Computing*, Vol. 11, No. 3, Sep 2004, pp. 279-302.
- [17] J. Meseguer, "Conditional rewriting logic as a unified model of concurrency", in *Theoretical Computer Science*, Vol. 96(1), 1992, pp. 73-155.
- [18] M. Clavel, F. Durán, S. Eker, P. Lincoln, N. Martí-Oliet, J. Meseguer and C. Talcott, "Maude manual (version 2.2)", Internal Report, SRI International, December 2007.
- [19] S. Eker, J. Meseguer and A. Sridharanarayanan, "The Maude LTL model checker", in *Proceedings of the 4th International Workshop on Rewriting Logic and Its Applications (WRLA)*, *Electronic Notes in Theoretical Computer Science*, Vol. 71, 2002.
- [20] J. De Lara and H. Vangheluwe, "Meta-Modelling and Graph Grammars for Multi-Paradigm Modelling in AToM³", in *Software and Systems Modelling, Special Section on Graph Transformations and Visual Modeling Techniques*, Vol. 3, 2004, pp. 194–209.
- [21] R. Bardohl, H. Ehrig, J. De Lara and G. Taentzer, "Integrating Meta Modelling with Graph Transformation for Efficient Visual Language Definition and Model Manipulation", in *Wermelinger, M., Margaria-Steffen, T. (eds.) FASE 2004. LNCS Springer, Heidelberg*, Vol. 2984, 2004, pp. 214–228.

Particle swarm optimization and method of moments for modeling and optimization of microstrip antennas

Tarek Fortaki and Mounir Amir

Electronics Department
University of Batna
Batna, Algeria
t_fortaki@yahoo.fr

Siham Benkouda and Abdelkrim Belhedri

Electronics Department
University of Frères Mentouri – Constantine 1
Constantine, Algeria
s_benkouda@yahoo.fr

Abstract— This paper introduces a novel technique for efficiently combining Particle Swarm Optimization (PSO) with method of moments (MOM) for computing the resonant frequency and bandwidth of rectangular microstrip antenna. In this technique, the problem is formulated in terms of an integral equation which is the kernel of a dyadic Green's function. After this PSO which will be introduced to determine antenna parameters by optimizing the impedance matrix $|Z|$. The resonant frequency results obtained by using (PSO/MOM) algorithm are in very good agreement with the experimental results available in the literature. The computation time is greatly reduced as compared to the classical MOM.

Keywords-component—particle swarm optimization (PSO); method of moments (MOM); microstrip antenna; modeling; optimization.

I. INTRODUCTION

Due to their many attractive features, microstrip antennas have drawn the attention of researchers over the past decades. Microstrip antennas are used in an increasing number of applications, ranging from biomedical diagnosis to wireless communication. Research on microstrip antenna in the 21st aims at size reduction, high gain, resonant frequency, wide bandwidth, multiple functionality, and system-level integration [1]. Several methods are available in the literature for computing the resonant frequency of patch antennas. These methods can generally be divided into two groups: simple analytical methods and rigorous numerical methods. For rigorous methods like the moment method (MOM) the exact mathematical formulations involve extensive numerical procedures. This technique needs an important time for calculation. A new method, the evolutionary algorithms, in particular the genetic algorithms (GA), have been widely used in electromagnetic applications in the last years [2]. More recently, a new stochastic optimization technique has rapidly gained popularity in the Electromagnetic Community: the Particle Swarm Optimization (PSO). PSO can be understood as a modeling via an analogy similar to the social activities of a bird flock, or a bee swarm. The PSO is a powerful technique that greatly simplifies the optimization process compared to ANN and

GA [3]. In this paper, the combination of Particle Swarm Optimization (PSO) and moment method (MOM) is presented for the calculation of the resonant frequency and bandwidth. This combination is intended to reduce the computation time, and at the same time keep the quality of the results obtained by the moment method. The same algorithm can be also used for the optimization of geometrical parameters of microstrip antennas.

II. ANALYSIS METHOD

The problem to be solved is illustrated in figure 1. We have a rectangular patch antenna with dimensions $(a \times b)$, is printed on a dielectric substrate isotropic of thickness d , is characterized by the free space permeability μ_0 and a permittivity ϵ . The ambient medium is air with constitutive parameters μ_0 and ϵ_0 . Assuming an $e^{i\omega t}$ time variations and starting from Maxwell's equations in the Fourier transform domain, we can show that the transverse fields inside the j layer ($Z_{j-1} < Z < Z_j$) can be written in terms of the longitudinal components \vec{E}_z and \vec{H}_z as [4-5].

$$\tilde{\mathbf{E}}(k_s, z) = \begin{bmatrix} \tilde{E}_x(k_s, z) \\ \tilde{E}_y(k_s, z) \end{bmatrix} = \bar{\mathbf{F}}(k_s) \cdot \begin{bmatrix} \frac{1}{k_s} \frac{\partial \tilde{E}_z(k_s, z)}{\partial z} \\ \frac{\omega \mu_0}{k_s} \tilde{H}_z(k_s, z) \end{bmatrix}$$

$$\tilde{\mathbf{E}}(k_s, z) = \bar{\mathbf{F}}(k_s) \cdot \mathbf{e}(k_s, z) \quad (1)$$

$$\tilde{\mathbf{H}}(k_s, z) = \begin{bmatrix} \tilde{H}_y(k_s, z) \\ -\tilde{H}_x(k_s, z) \end{bmatrix} = \bar{\mathbf{F}}(k_s) \cdot \begin{bmatrix} \frac{\omega \varepsilon_j}{k_s} \tilde{H}_z(k_s, z) \\ \frac{1}{k_s} \frac{\partial \tilde{H}_z(k_s, z)}{\partial z} \end{bmatrix}$$

$$\tilde{\mathbf{H}}(k_s, z) = \bar{\mathbf{F}}(k_s) \cdot \mathbf{h}(k_s, z) \quad (2)$$

Where \mathbf{e} and \mathbf{h} are, respectively, the transverse electric and magnetic fields in the (TM, TE) representation, and

$$\bar{\mathbf{F}}(k_s) = \frac{1}{k_s} \begin{bmatrix} k_x & k_y \\ k_y & -k_x \end{bmatrix} \quad (3)$$

With : $k_s^2 = k_x^2 + k_y^2$

Substituting the expressions of \tilde{E}_z and \tilde{H}_z [4-6] into (1) and (2), we get

$$\tilde{E}_z = A e^{-ik_z z} + B e^{ik_z z} \quad (4)$$

$$\tilde{H}_z = \bar{g}(k_s) \cdot [A e^{-ik_z z} - B e^{ik_z z}] \quad (5)$$

In (4) and (5), A and B are two-component unknown vectors and

$$\bar{g}(k_s) = \text{diag} \left[\frac{\omega \varepsilon}{k_z}, \frac{k_z}{\omega \mu} \right] \quad (6)$$

Writing (4) and (5) in the planes $z = z_{j-1}$ and $z = z_j$, and by eliminating the unknowns A and B, we obtain the matrix form

$$\begin{bmatrix} \mathbf{e}(k_s, z_j^-) \\ \mathbf{h}(k_s, z_j^-) \end{bmatrix} = \bar{\mathbf{T}}_j \cdot \begin{bmatrix} \mathbf{e}(k_s, z_{j-1}^+) \\ \mathbf{h}(k_s, z_{j-1}^+) \end{bmatrix} \quad (7)$$

With

$$\bar{\mathbf{T}}_j = \begin{bmatrix} \bar{\mathbf{T}}_j^{11} & \bar{\mathbf{T}}_j^{12} \\ \bar{\mathbf{T}}_j^{21} & \bar{\mathbf{T}}_j^{22} \end{bmatrix}$$

$$\bar{\mathbf{T}}_j = \begin{bmatrix} \cos(k_{z_j} d_j) & -i \bar{\mathbf{g}}^{-1} \cdot \sin(k_{z_j} d_j) \\ -i \bar{\mathbf{g}} \cdot \sin(k_{z_j} d_j) & \cos(k_{z_j} d_j) \end{bmatrix} \quad (8)$$

Which combines e and h on both sides of the j^{th} layer as input and output quantities. The matrix $\bar{\mathbf{T}}_j$ is the matrix representation of the j^{th} layer in the (TM, TE) representation. The boundary conditions for the considered structure presented by (fig.1) in the spectral domain

$$\bar{\mathbf{e}}_1(k_s, z_0^+) = \bar{\mathbf{0}} \quad (9)$$

$$\begin{bmatrix} \bar{\mathbf{e}}_2(k_s, z_1^+) \\ \bar{\mathbf{h}}_2(k_s, z_1^+) \end{bmatrix} = \bar{\mathbf{T}}_1 \cdot \begin{bmatrix} \bar{\mathbf{e}}_1(k_s, z_0^+) \\ \bar{\mathbf{h}}_1(k_s, z_0^+) \end{bmatrix} - \begin{bmatrix} 0 \\ \bar{\mathbf{J}}(z_1) \end{bmatrix} \quad (10)$$

$$\bar{\mathbf{h}}_2(k_s, z_1^+) = \bar{\mathbf{g}}_0(k_s) \cdot \bar{\mathbf{e}}_2(k_s, z_1^+) \quad (11)$$

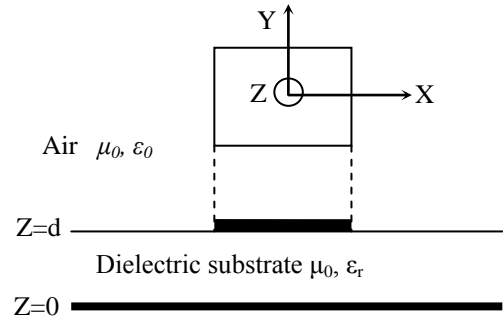


Fig. 1 .Geometrical structure of a rectangular microstrip patch.

The transformed components of the tangential electric field are expressed as function of the transformed current density components on the patch, as

$$\begin{bmatrix} \tilde{E}_x \\ \tilde{E}_y \end{bmatrix} = \begin{bmatrix} G_{xx} & G_{xy} \\ G_{yx} & G_{yy} \end{bmatrix} \cdot \begin{bmatrix} \tilde{J}_x \\ \tilde{J}_y \end{bmatrix} \quad (12)$$

The Galerkin moment method is implemented in the Fourier transform domain to reduce the integral equation to a matrix equation. The surface current \mathbf{J} on the patch is expanded into a finite series of known basis functions J_{xm} and J_{ym}

$$\mathbf{J} = \sum_{n=1}^N a_n \begin{bmatrix} J_{xn} \\ 0 \end{bmatrix} + \sum_{m=1}^M a_m \begin{bmatrix} 0 \\ J_{ym} \end{bmatrix} \quad (13)$$

Where n and m are the mode expansion coefficients to be sought. Substituting the vector Fourier transforms. Next, the resulting equation is tested by the same set of basis functions that was used in the expansion of the patch current. Thus, the integral equation is discredited into the following matrix equation:

$$\begin{bmatrix} (\bar{Z}_{kn}^1)_{N \times N} & (\bar{Z}_{km}^2)_{N \times M} \\ (\bar{Z}_{lm}^3)_{M \times N} & (\bar{Z}_{ln}^4)_{M \times M} \end{bmatrix} \begin{bmatrix} (a_n)_{N \times 1} \\ (b_m)_{M \times 1} \end{bmatrix} = \begin{bmatrix} \bar{0} \\ \bar{0} \end{bmatrix} \quad (14)$$

$$Z_{kn}^1 = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \tilde{J}_{xk}(-k_x, -k_y) G_{xx} \tilde{J}_{xm}(k_x, k_y) dk_x dk_y$$

$$Z_{km}^2 = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \tilde{J}_{xk}(-k_x, -k_y) G_{xy} \tilde{J}_{ym}(k_x, k_y) dk_x dk_y$$

$$Z_{lm}^3 = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \tilde{J}_{yl}(-k_x, -k_y) G_{yx} \tilde{J}_{xm}(k_x, k_y) dk_x dk_y$$

$$Z_{ln}^4 = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \tilde{J}_{yl}(-k_x, -k_y) G_{yy} \tilde{J}_{ym}(k_x, k_y) dk_x dk_y$$

The system of linear equations given in Eq. 14 has non-trivial solutions when

$$\det[Z(\omega)] = 0 \quad (15)$$

Equation 15 is an eigen equation for ω , from which the characteristics of the stacked structure of Figure1 can be obtained. In fact, let $\omega=2\pi(f_r+if_i)$ be the complex root of Eq. 15. In that case, the quantity f_r stands for the resonant frequency, the quantity $BW = 2f_i/f_r$ stands for the bandwidth. In our case eq. 15 represent the fitness function that will be optimized by using PSO algorithm.

III. PARTICLE SWARM OPTIMIZATION

The Particle Swarm Optimization is an evolutionary optimization technique, which is a local to global search method based on particle search. It follows the optimization process by means of local best (pbest), global best (gbest), particle displacement and particle velocity. In this paper, all these features have been applied on the determination of the resonant

frequency (f_r) and bandwidth (BW), function of geometric parameters and dielectric constants of rectangular microstrip antenna. The global function approximation capability and generalization capability of Particle Swarm Optimization in the modeling of microstrip antenna have also been studied. This method is used for particles from landing on any solution instead of just the best solution. This is where the social aspect of mind and intelligence comes into play. The particles are considered to move through co-ordinates of N-dimensional space. When particle moves, it sends its co-ordinates to a function and measures its "fitness" value, close to a best solution for the problem. The evolutions of particles, guided only by the best solution, tend to be regulated by behavior of the neighbors. In the simplest form, the position and velocity 'v' of each particle are represented by the following equations considering *pbest* rather than *gbest* as the best position of the particle referred to the neighbors. The particle velocity is expressed as [7].

$$v_i(k) = wv_i(k-1) + C_1r_1 \times (pbest_i - S_i(k-1)) + C_2r_2 \times (gbest - S_i(k-1)) \quad (16)$$

where, $v_i(k)$: velocity of agent i at iteration k ,
 w : weighting function,
 C_1, C_2 : weighting factor,
 r_1, r_2 : random number between 0 and 1,
 $S_i(k)$: current position of agent i at iteration k ,
 $pbest_i$: pbest of agent i ,
 $gbest$: gbest of the group.

Using the above equation, a certain velocity which gradually gets close to *pbest* and *gbest* can be calculated. The current position (searching point in the solution space) can be modified by the following equation:

$$S_i(k) = S_i(k-1) + v_i(k)$$

The basic program flow of PSO is depicted in a flowchart as shown in Fig. 2. As mentioned, the objective of the optimization is the optimization of the matrix impedance (Z) in Eq. 15.

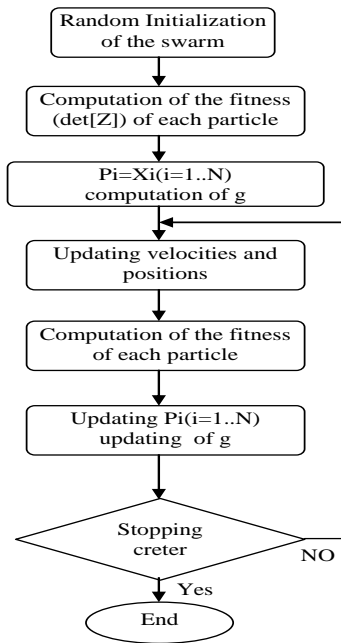


Fig. 2. Flowchart of the PSO/MOM algorithm.

IV. RESULTS

The resonant frequencies calculated by using PSO/MOM algorithm presented in this paper for electrically thin and thick

TABLE 1. COMPARISON OF THE CALCULATED RESONANT FREQUENCY WITH MEASURED AND CALCULATED DATA, FOR A RECTANGULAR MICROSTRIP ANTENNA.

a (cm)	b (cm)	d (cm)	ϵ_r	Resonant frequencies (GHz)	
				Measured	PSO/MOM
1.900	2.290	0.1590	2.32	4.104*	4.1053
5.700	3.800	0.3175	2.33	2.310 ⁺	2.3278
4.550	3.050	0.3175	2.33	2.890 ⁺	2.6225
2.950	1.950	0.3175	2.33	4.240 ⁺	4.1179
1.950	1.300	0.3175	2.33	5.840 ⁺	5.5534
1.700	1.100	0.1375	2.33	6.800 ⁺	6.5377
1.400	0.900	0.3175	2.33	7.700 ⁺	7.5360
1.200	0.800	0.3175	2.33	8.270 ⁺	8.1060
1.050	0.700	0.3175	2.33	9.140 ⁺	9.4870
1.700	1.100	0.9525	2.33	4.730 ⁺	4.5026
1.700	1.100	0.1524	2.33	7.870 ⁺	7.5110
4.100	4.140	0.1524	2.50	2.228 ^Δ	2.2014
6.858	4.140	0.1524	2.50	2.200 ^Δ	2.1148
10.80	4.140	0.1524	2.50	2.181 ^Δ	2.0173
2.000	2.500	0.0790	2.22	3.970 ^Γ	3.7141
1.120	1.200	0.2420	2.55	7.050 ^Γ	7.0224
0.790	1.255	0.4000	2.55	7.134 ^Γ	7.1187

* This frequency reported by Fortaki et al. [9]

+ These frequencies measured by Chang et al. [10].

Δ These frequencies measured by Carver. [11].

Γ Measured by Kara [12, 13].

V. CONCLUSIONS

In this paper, an efficient method for the integration of Particle Swarm Optimization (PSO) with the method of moments (MOM) for microstrip antenna modeling has been presented. PSO/MOM was applied successfully for determination of resonant frequency and bandwidth of a rectangular patch antenna. The calculated results have been compared with measured one available in the literature and excellent agreement has been found. Better accuracy with respect to the previous conventional methods and natural selection method (like PSO) is obtained. Since the method presented in this paper have good accuracy, require no tremendous computational effort, they can be very useful for the development of fast CAD algorithms.

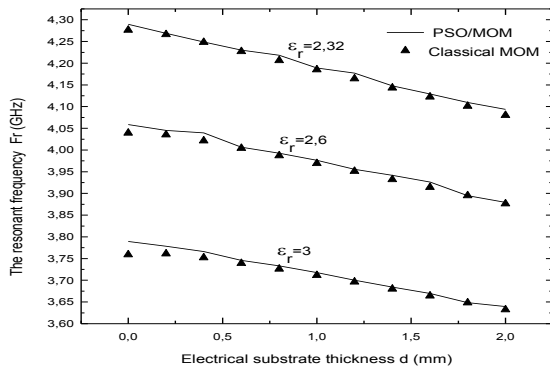


Fig. 3. The resonant frequency versus electrical substrate thickness.

substrates are depicted in Fig. 3. It is evident from Fig 3 that the resonant frequency indeed decreases as the antennas become electrically thicker as has been shown in previously published results in literature [9]

rectangular MSAs are listed and compared with measured results in Table I. The results of this method are in very good agreement with measurements. It can be very useful for the development of fast CAD algorithms.

Fig.3 and Figure 4 shows the resonant frequency and bandwidth as a function of substrate thickness for different values of dielectric constant $\epsilon_r = 2.32, 2.6$ and 3 of a rectangular patch antenna with dimensions $a = 1.9\text{cm}, b = 2.29\text{cm}$. From this graphs we see that the results obtained by PSO/MOM algorithms have the same behavior as those obtained by the method of moment.

It should be noted, that the time necessary for calculates parameters of the antenna by PSO/MOM is approximately 50s but for traditional moment method is between 5 and 20mn.

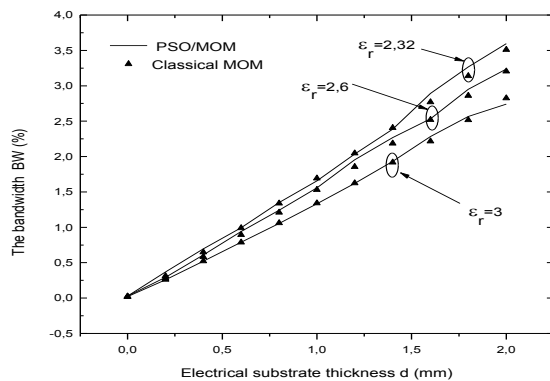


Fig. 4. The bandwidth versus electrical substrate thickness.

REFERENCES

- [1] D. K. Neag, S. Pattnaik, D.C. Panda, S.Devi, B.Khuntia and M. Dutta, "Design of a wideband microstrip antenna and the use of artificial neural networks in parameter calculation", IEEE Transactions on Antennas and Propagation, VOL. 45, NO. 3, JUNE 2005.
- [2] S. Sella, M. Mussetta, P. Pirinoli, R. E. Zich, , and L. Matekovits, "Some Insight Over New Variations of the Particle Swarm Optimization Method", IEEE Antennas And Wireless Propagation Letters, VOL. 5, 2006.
- [3] K. L. Chung and W. Y. Tam; "Particle Swarm Optimization of Wideband Patch Antennas" Microwave Conference, 2008. APMC 2008. Asia-Pacific; 16-20 Dec. 2008
- [4] J.S. Dahele and K.F. Lee, "Theory and experiment on microstrip antennas with air gap", IEE Proc Pt H 132 (1985), 455-460.
- [5] D. Guha, "Resonant frequency of circular microstrip antennas with and without air gaps", IEEE Trans on Antennas and Propagat 49 (2001), 55-59.
- [6] D. Guha, "Resonant frequency of circular microstrip antennas with and without air gaps", IEEE Trans Antennas Propag 49 (2001), 55-59.
- [7] V. S. Chintakindi, , S. S. Pattnaik, , O.P.Bajpai, S.Devi, "Resonant Frequency of Equilateral Triangular Microstrip Patch Antenna Using Particle Swarm Optimization Technique", Proceedings of International Conference on Microwave -2008
- [8] Y. Shi and Eberhart, "R.CA modified particle swarm optimizer", Proceedings of the IEEE International Conference on Evolutionary Computation., IEEE Press, Piscataway, NJ, 1998, pp.69-73
- [9] T. Fortaki, D. Khedrouche, F. Bouttot, A. Benghalia, "A numerically efficient full-wave analysis of a tunable rectangular microstrip patch", INT. J. Electronics, Vol. 91, No 1, (2004), 57-70.
- [10] E.Chang, Long S.A., Richards W.F., "An experimental investigation of electrically thick rectangular microstrip antennas", IEEE Trans. Antennas propagate. 34 (1986), 767-772.
- [11] K.R., Carver "Practical analytical techniques for the microstrip antenna", Proc. Workshop on printed Circuit antenna technology, New Mexico State University, Las Cruces, (Oct 1979), 7,1-7,20.
- [12] M. Kara. "The resonant frequency of rectangular microstrip antenna elements with various substrate thicknesses". Microwave and Optical Technology Lett, 11(1996), 55-59.
- [13] M. Kara. "Closed-form expressions for the resonant frequency of rectangular microstrip antenna elements with thick substrates". Microwave and Optical Technology Lett. 12 (1996). 131-136.

Simulation of Class D resonance inverter for Acoustics Energy Transfer applications

Huzaimah Husin

Faculty of Electronics & Computer Engineering
Universiti Teknikal Malaysia Melaka
Melaka, Malaysia
huzaimah@utem.edu.my

H. Hamidon

Faculty of Electronics & Computer Engineering
Universiti Teknikal Malaysia Melaka
Melaka, Malaysia
hamid@utem.edu.my

Shakir Saat

Faculty of Electronics & Computer Engineering
Universiti Teknikal Malaysia Melaka
Melaka, Malaysia
shakir@utem.edu.my

Y. Yusmarnita

Faculty of Electronics & Computer Engineering
Universiti Teknikal Malaysia Melaka
Melaka, Malaysia
yusmarnita@utem.edu.my

Abstract – Power conditioning stage of an acoustics energy transfer (AET) is a key step in determining the overall efficiency of the AET system. On the transmitter side, it needs to drive the device at the exact operating frequency meanwhile at the receiver side, it need to able to extract maximum power through interfacing with PZT transducer. This paper will study and simulate the Class D parallel-resonance inverter for the transmitter side of an AET system. The aimed of the circuit is to produce around 80 mW power at the operating frequency 416 kHz for implantable applications. The Proteus Software used as simulation platform with peripheral interface controller (PIC16F877A) as pulse width modulation signal generator. The pulse width modulation (PWM) used to generate switching signal for MOSFET IRF5852TR gate in the circuit.

Keywords—*Half-bridge Class D inverter; low power applications*

I. INTRODUCTION

In certain application areas, contactless energy transfer or wireless power transfer (WPT) is increasingly being considered to be viable alternative for the power supply of electronics. Consumer products employ WPT to charge batteries of mobile devices without having to use an adaptor [1]. It also used in industry in linear and planar actuators in which the wear of cable slabs and the parasitic force that exerted have to be minimized [2]. Lastly, WPT can also be applied to charge microbatteries in biomedical implants [3]. The contactless transmission of energy consists a number of established technologies such as inductive energy transfer (IET), capacitive energy transfer (CET), far-field electromagnetic coupling (EM) and optical coupling. A relatively new method is acoustic energy transfer (AET) that uses vibration or ultrasound waves to propagate energy without relying on electrical contact. AET is an effective method for powering implanted devices such as pacemakers, defibrillators, heart-assist devices and implanted insulin pumps. Some of the implanted devices are designed to serve monitoring purposes such as biosensors, glucose indicators. Even though IET has been receiving considerable attention lately, with the recent publications on systems delivering energy over distances up to 2 m at high efficiencies [4] and [5] but due to the magnetic coupling technique, IET

is not suitable for transferring the power across metal objects and can cause large eddy current losses [6],[7] and [8]. In order to overcome these limitations, CET is used since an electric field can penetrate through any metal shielding environment. The CET not only can transmit through metal and shielded body, but also has good anti-interference ability of magnetic field [7],[8],[9] and [10]. However, till recent, CET systems have only been used for very low power delivery applications [6], [8],[9] and [10]. CET is used far less often due to the limitation of distance that can be crossed with it. This is a direct consequence of the inverse proportionality of the capacitance with the distance, requiring high voltages and frequencies for the transfer of a certain amount of power.

Another principle for WPT is far-field EM or microwave energy transfer is seldom used. Instead of the nonradiative used in inductive and capacitive cases, a radiative EM field functions as the energy transfer medium. Rectification of these high-frequency waves at the pick-up unit can be achieved at high efficiency of 80% - 90% [11]. Generation of the microwaves, on the other hand, is much more difficult, particularly when a solid-state RF generator is used.

Optical energy transmission uses same principle as far-field EM and has low efficiency whereby 40% and 50% of energy is lost [10] and [12].

All the previously described methods rely on EM fields for the transfer of power. The radiative wave that is used in microwave and optical WPTs hints at an entirely different way of wireless conveying power. One just has to realize that such a wave does not have to be of an EM nature but that any type of wave can be used for this purpose. The choice of using sound wave or ultrasonic waves therefore seems a reasonable one.

II. ACOUSTICS ENERGY TRANSFER CONCEPT

Impracticality the use of battery and physically wired for the implantable devices inspires this research to be carried out. The application of ultrasound or vibration as the medium of energy transmission especially in situations where no EM fields are allowed, and high directionality of the power transfer in combination with small system dimensions is required [10] and [12].

A. AET System

The basic structure of AET system is shown in Fig. 1. This system consists of transmitter unit and receiver unit. On the transmitter unit, external transducer usually PZT transducer will be attached to the skin surface and facing an implanted transducer at the receiver unit.

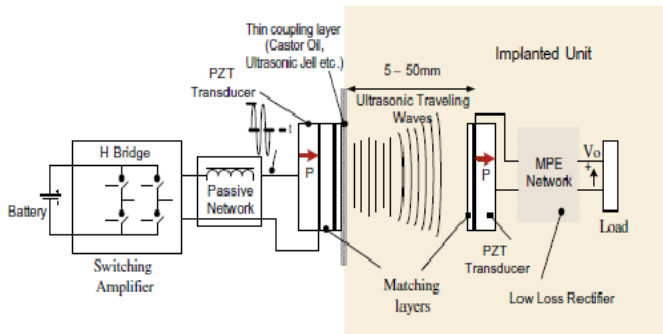


Fig. 1. Acoustic Energy Transfer system [3]

An electrical power source, a DC will energizes the transmitter transducer that converts the electrical energy to vibration or acoustics pressure waves. The acoustic waves will propagates and carry the energy through the tissue toward an implanted receiver transducer that positioned within the radiation lobe of the transmitter. A receiving unit functions for the inverse process of converting the vibration or motion caused by the waves into electrical energy. A rectifier will rectify that particular electrical energy and filters the output voltage of the receiving transducer. The usable steady dc voltage drives a load. In order to minimize the inconvenience caused to the patient as well as to ensure a close fit to the body (which is required for good acoustic coupling), the

device should be light in weight and thin so that its center of gravity is as close as possible to the surface of the skin.

B. Power Conditioning

The overall efficiency of AET system will be determined by the power conditioning stage in the circuit. It is requires to drive the device at the exact operating frequency without exciting harmonic modes at the transmitter unit. On the receiver unit, the circuit should interface with the transducer so as to extract maximum power. The power conditioning circuit on the both sides must have efficiency >80% as they affect the overall efficiency of the energy transfer. This paper will focus on powering the transmitter unit of AET system. The CLASS D inverter is one of the high-frequency high-efficiency resonant power sources, which has been applied to dc/dc resonant converters, radio transmitters, and electronic ballasts for fluorescent lamps[13] and [14]. Its high dc/ac power conversion efficiency is achieved by the zero-current switching (ZCS), which enables its operation at frequency of several hundred kilohertz. Furthermore, this resonant inverter with sinusoidal waveforms achieves low switching losses due to the phase displacement between the voltage and current through the transistors [15].

C. Power Requirement

The key to settling the impracticality the use of battery and physically wired for the implantable devices problem is by having the continuous supply of a stable power source. In most cases, the devices has to be replaced just because of the battery is running out inside the device. Therefore, it is the battery that determines the longevity of the devices. Although the requirement of each device is different, the power that needed generally falls in the level of μW – mW as in Table 1.

TABLE 1. POWER REQUIREMENT FOR IMPLANTABLE DEVICES [16]

Implanted devices	Typical power requirement
Pacemaker	30 μ -100 μ W
Cardiac defibrillator	30 μ -100 μ W
Neurological simulator	30 μ to several mW
Drug pump	100 μ to 2mW
Cochlear implant	10mW

III. THEORETICAL RESULTS

In order to design the transmitter side, which focuses on half-bridge Class D parallel-resonance inverter, the theoretical value of each components is obtained through calculation. The equations related were explained in details in [17]. The calculation based on the standard circuit shown in Fig. 2.

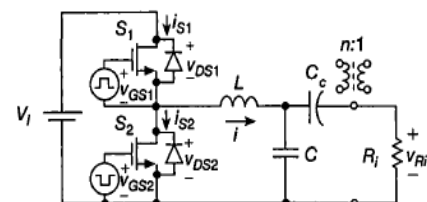


Fig. 2. Standard Circuit of the Class D parallel resonant inverter [17]

A. Principle of operation

A standard circuit of half-bridge Class D voltage-source parallel-resonant inverter (PRI) is shown in Fig.2. It consists of two switches S_1 and S_2 , a resonant inductor L , a resonant capacitor C , and a DC-blocking capacitor C_c . Resistance R_i represents a load to which the AC power is to be delivered and is connected in parallel with the resonant capacitor C . The average voltage across capacitor C_c is equal to $V_1/2$. The two bidirectional switches S_1 and S_2 and the DC input voltage source V_1 form a square-wave voltage source that drives the resonant circuit L - C - R_i . Switches S_1 and S_2 are alternately turned ON and OFF at the switching frequency $f = \omega/2\pi$. Because of the turn-off delay time of power MOSFETs, the duty cycle of drive voltages V_{GS1} and V_{GS2} should be slightly less than 50% to avoid cross conduction.

B. Assumptions of analysis

The analysis of the inverter in Fig. 2 assumes

- i. Each switch is a resistance r_{DS} when “on” and an open circuit when “off”.
- ii. Switching losses are neglected.
- iii. The loaded quality factor Q_L of the resonant circuit is high enough so that the currents through inductance L , capacitance C , and resistance R_i is sinusoidal.
- iv. The coupling capacitance C_c is high enough so that its AC voltage ripple is negligible.
- v. The output capacitances of MOSFETs are neglected.

C. Design parameters

- i. Input voltage, $V_1 = 3.6\text{Vdc}$
- ii. Output power, $P_{Ri} = 80\text{ mW}$
- iii. Resonant frequency, $f_r = 416\text{ kHz}$.
- iv. Quality factor, $Q_L = 2.5$

D. Equations

In order to get theoretical value of components, some equations used as stated as below.

Assume a typical value of the inverter efficiency $\eta_I = 95\%$, some relevant equations as below:

DC supply power is

$$P_1 = \frac{P_{Ri}}{\eta_I} \text{ W} \quad (1)$$

DC supply current is

$$I_1 = \frac{P_1}{V_1} \text{ A} \quad (2)$$

Assuming $f = f_r = 416\text{ kHz}$ at full power, the corner frequency is

$$f_o = \frac{f_r}{\sqrt{1 - \frac{1}{Q_L}}} \text{ Hz} \quad (3)$$

The AC load resistance

$$R_i = \frac{V_{Ri}^2}{P_{Ri}} = \frac{2V_1^2 \eta_I^2}{\pi^2 P_{Ri} \left(\left[1 - \left(\frac{\omega}{\omega_o} \right)^2 \right]^2 + \left[\frac{1}{Q_L} \left(\frac{\omega}{\omega_o} \right) \right]^2 \right)} \Omega \quad (4)$$

The characteristic impedance can be obtained as

$$Z_o = \frac{R_i}{Q_L} \Omega \quad (5)$$

Thus, the elements of resonant circuits are

$$L = \frac{Z_o}{\omega_o} \text{ Henry} \quad (6)$$

and

$$C = \frac{1}{\omega_o Z_o} \text{ Farad} \quad (7)$$

The maximum value of the switch peak current is

$$I_{m(max)} = I_{SM(max)} = \frac{2V_1 \sqrt{Q^2 L + 1}}{\pi Z_o} \text{ A} \quad (8)$$

The voltage stresses on the resonant components are

$$V_{Cm(max)} = \frac{2V_1 Q_L}{\pi} \text{ V}, \quad V_{Lm(max)} = \frac{2V_1 \sqrt{Q^2 L + 1}}{\pi} \text{ V} \quad (9)$$

As the load is in parallel with resonant capacitor as shown in Fig. 2, the output voltage at the load can be obtained as

$$V_o = V_{cm} V \tag{10}$$

As the aim of this paper is to produce output power at R_i , the equation below use to calculate the output power required.

$$P_{Ri} = \frac{V_{o_{rms}}^2}{R_i} \tag{11}$$

The selection of IRF5852TR as a switching MOSFET due to suitable voltage rating and power dissipation regards to design specifications.

Using equations (1)-(11), the theoretical value of each component and parameter were calculated and tabulated in Table 2.

TABLE 2. CALCULATION VALUE FOR EACH COMPONENT AND PARAMETERS FOR INVERTER

Inverter Parameters	Symbol	Value
Dc Supply Power	P_1	84.21 mW
Dc Supply Current	I_1	23.4 mA
Corner frequency	f_o	453.89 kHz
AC Load Resistance	R_i	185.28 Ω
Impedance	Z_o	74.11 Ω
Resonant Inductor	L	25.9 μ H
Resonant Capacitor	C	4.73 nF
Switch Peak Current	$I_{m(max)}$	83.2 mA
Voltage at resonant Capacitor	$V_{Cm(max)}$	5.73 V
Voltage at resonant Inductor	$V_{Lm(max)}$	6.17 V
Output power gained	P_{Ri}	88.6 mW

IV. SIMULATION RESULTS

In order to verify the designed model, simulation through Proteus Software where the inputs signal PWM was generated by peripheral interface controller (PIC16F877A). The coding for the PWM was built using mikroC PRO for PIC software.

A. The design of power amplifier Circuit

The circuit designed by applying MOSFET IRF5852TR as a switch that turns on and off alternately at the switching frequency, $f = \frac{\omega}{2\pi}$. The square wave signal as a driver for those MOSFETs was generated by PIC and

feed in to the gate of the MOSFET. The designed circuit is shown in Fig. 3.

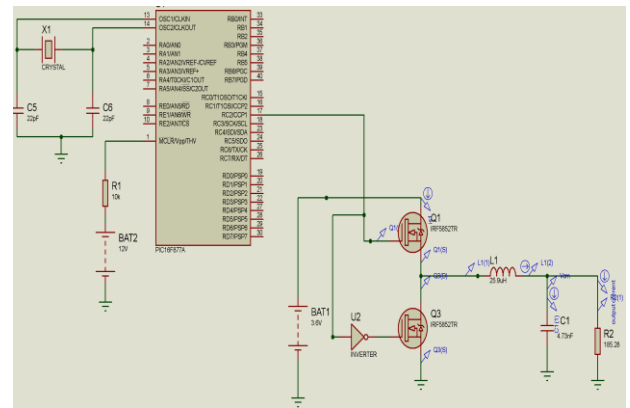


Fig. 3. The schematic of Class D Parallel-resonant inverter simulation

Proteus simulation has been undertaken to analyze the functionality of the designed circuit. The components value selection was based on the previous calculation as tabulated in Table 2.

B. The Gate Signal for MOSFET

The generation of PWM was done by PIC16F877A with the coding simulation through mikroC PRO for PIC software. The generation of 5Vp square wave is produced by PIC with the resonant frequency 416 kHz is successfully obtained and shown in Fig. 4. A 5Vp square wave input that drives the resonant circuit $L-C-R_i$ is connected to the gate of S_1 meanwhile the inverted is connected to the gate of S_2 . The inversion is done in order to fulfill the out-of-phase condition between S_1 and S_2 and shown in Fig. 5.

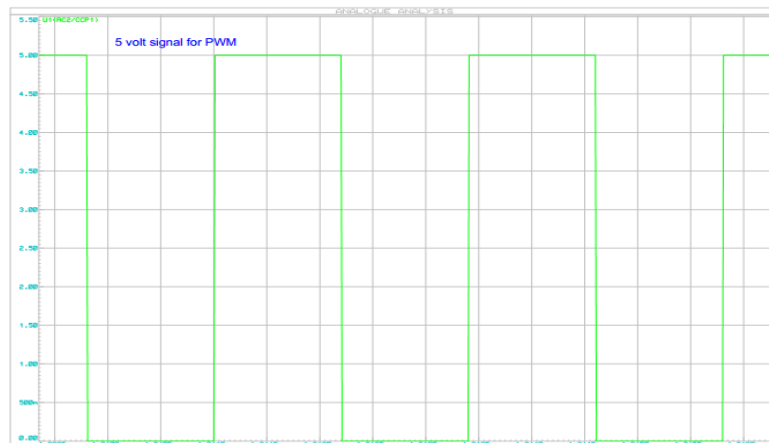


Fig. 4. Simulated gate signal generated by PIC

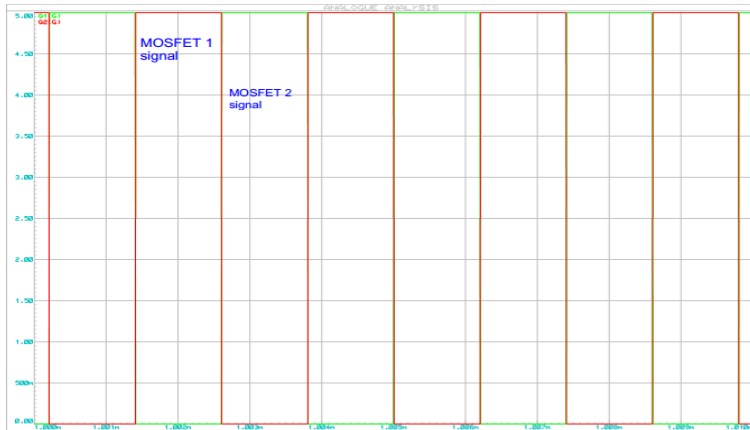


Fig. 5. Simulated gate signal for MOSFET

C. The Output Power of the circuit

This paper concentrated on producing low power output for implantable devices that uses acoustic as medium of propagation. The output power calculated by using equation (10) and (11). Fig. 6 shows the simulated output voltage of the inverter.

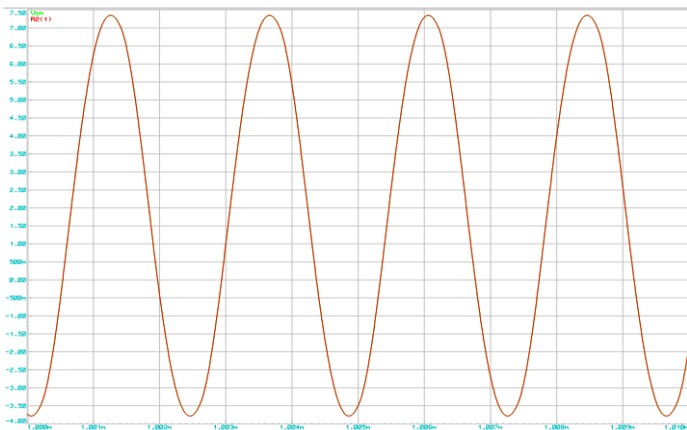


Fig.6 . Simulated output voltage of the inverter

From the graph, the peak-peak output voltage is 11.2 V; meanwhile the peak output voltage is 5.6 V. The fully sinusoidal waveform is successfully obtained at the output voltage thus consistent with the purpose of inverter that converted a DC voltage to an AC voltage. The output power was measured at the point of AC load resistor, R_i . The waveform similarity also can be founded between V_{Cm} and V_o due to their parallel arrangement. From calculation, $V_o = V_{Cm} = 5.73$ V, meanwhile through simulation, $V_o = V_{Cm} = 5.6$ V. There is slightly different since the simulation might take into account some parasitic resistance of the components that affected the power. Using equation (11), this design obtained 84.63 mW as output power compared to 88.6 mW as in the calculation.

V. CONCLUSION

In this paper, the requirement of powering AET system at the transmitter side is studied and the circuit of power amplifier is designed. Various important components value were calculated and tabulated based on established equations. To validate the theoretical results, the simulation was undertaken and the results shows that the output voltage consistent with the Class D parallel-resonant inverter characteristics. Observed waveforms and values in the circuit simulations showed good agreement with the calculated ones. Future work on this design to consider various parameters such as load, quality factor and MOSFET selection will be done.

ACKNOWLEDGMENT

The author would like to express the appreciation to Universiti Teknikal Malaysia Melaka (UTeM) and Ministry Of Education Malaysia for funding this research work under RAGS/1/2014/TK03/FKEKK/B00062 grant.

REFERENCES

- [1] C. L. W. Sonntag, J. L. Duarte, and a. J. M. Pemen, "Load position detection and validation on variable-phase contactless energy transfer desktops," in *2009 IEEE Energy Conversion Congress Expo*, San Jose, California, 2009, pp. 1818–1825.
- [2] S. Hussmann and P. a. Hu, "A microcomputer controlled ICPT power pick-up and its EMC considerations for moving sensor applications," in *Proceedings of International Conference on Power System Technology.*, vol. 2, Kunming, China, 2002, pp. 1011–1015.
- [3] S. Ozeri and D. Shmilovitz, "Ultrasonic transcutaneous energy transfer for powering implanted devices.," in *Ultrasonics*, May 2010, vol. 50, no. 6, pp. 556–66.
- [4] A. Karalis, J. D. Joannopoulos, and M. Soljačić, "Efficient wireless non-radiative mid-range energy transfer," in *Ann. Phys. (N. Y.)*, 2008, vol. 323, pp. 34–48.
- [5] A. Kurs, A. Karalis, R. Moffatt, J. D. Joannopoulos, P. Fisher, and M. Soljagic, "Wireless power transfer via strongly coupled magnetic resonances.," in *Science*, 2007, vol. 317, no. July, pp. 83–86.
- [6] M. P. Theodoridis, "Effective capacitive power transfer," *IEEE Transaction on Power Electronics*, 2012, vol. 27, no. 12, pp. 4906–4913.
- [7] C. Liu, A. P. Hu, and M. Budhia, "A generalized coupling model for Capacitive Power Transfer systems," in *Annual*

- Conference of the IEEE Industrial Electronics Society*, Glendale, AZ, U.S.A., 2010, vol. 27, no. 12, pp. 274–279.
- [8] C. Y. Xia, C. W. Li, and J. Zhang, “Analysis of power transfer characteristic of capacitive power transfer system and inductively coupled power transfer system,” in *Proceeding . 2011 International Conference on Mechatronic Science. Electric Engineering and Computer*, Jilin, China, 2011, pp. 1281–1285.
- [9] M. Kline, I. Izyumin, B. Boser, and S. Sanders, “Capacitive power transfer for contactless charging,” in *IEEE Applied Power Electronics Conference Expo*, Forth Worth, Tx, 2011, pp. 1398–1404.
- [10] T. Zaid, S. Saat, Y. Yusop, and N. Jamal, “Contactless energy transfer using acoustic approach - A review,” in *IEEE 2014 International Conference on Computer, Communication and Control Technology*, Langkawi, Malaysia, 2014, pp. 376–381.
- [11] J. O. McSpadden and J. C. Mankins, “Space solar power programs and microwave wireless power transmission technology,” in *IEEE Microwave Magazine*, December, 2002, vol. 3.
- [12] M. G. L. Roes, S. Member, J. L. Duarte, M. A. M. Hendrix, E. A. Lomonova, and S. Member, “Acoustic Energy Transfer : A Review,” in *IEEE Transactions On Industrial Electronics*, Jan, 2013, Vol. 60, No. 1, pp. 242–248.
- [13] A. Ekbote and D. S. Zinger, “Comparison of class e and half bridge inverters for use in electronic ballasts,” in *Industry Applications Conference, 2006*, Tampa, FL, 2006, vol. 5, pp. 2198–2201.
- [14] H. Koizumi, K. Kurokawa, and S. Mori, “Analysis of Class D Inverter With Irregular,” in *IEEE International Symposium on Circuits and Systems*, Vancouver, Canada, 2004 vol. 53, no. 3, pp. 677–687.
- [15] C. Brañas, F. J. Azcondo, and R. Casanueva, “A generalized study of multiphase parallel resonant inverters for high-power applications,” in *IEEE Transactions on Circuits and Systems*. 2008, *1 Regular Paper*, vol. 55, no. 7, pp. 2128–2138.
- [16] X. Wei and J. Liu, “Power sources and electrical recharging strategies for implantable medical devices,” in *Front. Energy Power Eng. China*, 2008, vol. 2, no. 1, pp. 1–13.
- [17] Marian K. Kazimierczuk, “Inverter,” in *Resonant Power Converters*, 2nd ed., New Jersey, John Wiley & Sons, 2010, ch 7, sec. 7.2-7.3, p. 193–217.

Symbolic Modeling Approach in Verification and Testing

Oleksandr Letychevskiy

Department of Theory of Digital Automatic Machines
Glushkov Institute of Cybernetics of National Academy of Sciences
Kyiv, Ukraine
lit@iss.org.ua

Abstract—The paper outlines a symbolic modeling approach developed in Glushkov Institute of Cybernetics and applied in verification and model-based testing. This method is the result of 10 years of experience in a large amount of industrial projects in different subject domains. The models in this approach are presented as UCM (Use Case Maps) notation composed with basic protocols formal language. Symbolic modelling is used in verification of requirements and models of programs. It is also intended for creation of test suits and further test execution.

Keywords—symbolic modeling, symbolic execution, model-based testing, verification of requirements, predicate transformers

I. INTRODUCTION

The paper describes a symbolic modeling approach developed in Glushkov Institute of Cybernetics. Usage of symbolic modeling is the result of over 10 years of experience in the industrial application of formal methods in software development process especially for the requirements gathering and testing stages. It was deployed in a large number of projects at Motorola and Uniquesoft LLC in various subject domains, from telecommunications to networking, microprocessor programming, and automotive systems. The structure of the paper is the following.

In section 2 we consider related works dedicated to usage and development of deductive technologies and symbolic approach in software industry especially verification and testing. Section 3 describes the formal specifications implemented in UCM (Use Case Maps) notation and basic protocols language that are the input of our symbolic modeling method. Section 4 outlines the theoretical basics and semantics of symbolic modeling methods especially theory of predicate transformers developed in Glushkov Institute of cybernetics. Sections 5 - 7 describe usage of symbolic modeling in verification of requirements, test generation and test execution.

II. SYMBOLIC METHODS IN VERIFICATION AND TESTING

Usage of symbolic-based formal methods has become relevant due to the growth in complexity of the designs common in hardware and software industries. The term “symbolic” is closely related in meaning to “algebraic” where manipulating with mathematical expressions is anticipated. The methods of symbolic execution or symbolic modelling were invented in 70-th [1] but they became popular only now due to essential success in solving and proving techniques development.

SMT (Satisfiability Modulo Theory) solving approach was invented as a trend of model checking technique. Model checking [2] is an exhaustive exploration of the states and transitions of the mathematical model. Term “symbolic model checking” has been introduced by McMillan [3] where states of model are presented as formulas in some theory. It subsumes different SMT techniques and similar methods such as abstract interpretation [4], symbolic simulation [5], abstraction refinement [6], and others. Model checking is supported by a number of tools, such as SPIN or BLAST. The main objective of SMT-solving usage is a handling of states explosion problem while traversing the model states.

Usually SMT approach is applied in verification as a formal proof of properties of an abstract mathematical model of the system captured by the requirements or a program labeled by annotations. Examples of such mathematical models are finite state machines, labeled transition systems, Petri nets, or process algebras. There are two main approaches to establish properties of such models. In difference with exhaustive searching of the states set the other approach is deductive verification where the model or program specifications are presented as a set of assertions and the properties are established by theorem provers such as HOL, Z3, CVC or Isabelle.

Anyway, the detection of property true is defined by reachability of states that present this property. So the proving-based methods are complemented by the modelling features. The problem of reachability is unresolvable for systems with infinite number of states and methods of its detection were invented as heuristics for different classes of models.

So far the only few tools were presented as symbolic modeling for model-based testing. RT-tester [7] is commercial tool for symbolic trace generation with usage of SMT-solvers. Symbolic execution is also used for test generation in Symbolic Path Finder [8] for Java applications.

The main purpose of model-based testing is to obtain efficient coverage of model behavior. It is anticipated that the set of generated tests should cover the maximal set of states of models. There is a big variety of commercial MBT tools but they are not coping with huge or infinite number of states.

The more significant achievements in symbolic execution usage are the SAGA-project [9], KLOVER [10]. It implements symbolic modeling for detection of defects in code executing it symbolically.

The use of symbolic and deductive methods is difficult for engineers due to the complexity and unfamiliarity of formalization and verification methods. So one of the challenges aim at making symbolic techniques more amenable to industrial usage by front-ending mathematical techniques with familiar or easy to learn notations, including expressive graphical presentations.

III. PROBLEM STATEMENT AND INPUT MODELS

We consider two problems for application of symbolic modeling methods. The first is verification of models that present the program code or design models or requirements specifications.

Verification procedure is a searching of given property (or property violation) while model behavior. For example, it could be requirements specifications where the properties violations are inconsistency, incompleteness, safety violation or liveness. In a program code, it could be bounds violations, null pointer assignment or other possible errors.

The second is model-based testing where we have the model for test generation and problem is to generate the number of tests due to the customer request. Such request could contain coverage criterion that defines quality of test procedure as necessary covered states of model.

In both cases, we face with procedure of model formalization. It is the most time-consuming process because of manual creation of formal specifications from natural text or semi-formal descriptions, or other source of initial model presentation.

There is a number of widespread modeling notations such as UML2, SysML, ASM, Timed automaton (TASM). We consider the model of system as a composition of two notations. The base is UCM (Use Case Maps) notation standardized as part of URN (User Requirements Notation) in ITU-T recommendation (Z.151) [11] that provides a scenario-based approach to requirements specifications. UCM allows an easy and natural expression of sequences of events with synchronization and structure. The language of basic protocols developed at the Glushkov Institute of Cybernetics [12] extends UCM. It represents behavioral scenarios as reactions of a system to externally triggered events under some conditions. Such local behaviors are modeled by basic protocols that consist of three components:

- precondition defines the state of the environment of the system at the point when the basic protocol is applicable;
- process actions are presented as MSC (Message Sequence Chart) diagrams that show input and output signals and local actions;
- postcondition defines the change of the environment in response to the application of the basic protocol.

Pre- and postcondition are represented as formulae of the basic logic language. It supports attributes of numeric and symbolic types, arrays, lists, and functional data types. The following example of a basic protocol (Fig.1) is taken from the specification of a well-known telecommunication protocol.

The order and synchronization of basic protocols is defined by means of UCM diagrams in a graphical notation. UCM diagrams are oriented graphs with initial and final states. Nodes of the graph represent events in a system.

The basic protocols notation captures the atomic actions (responsibilities) of a UCM map. The UCM notation is a convenient tool for the description of parallel processes and

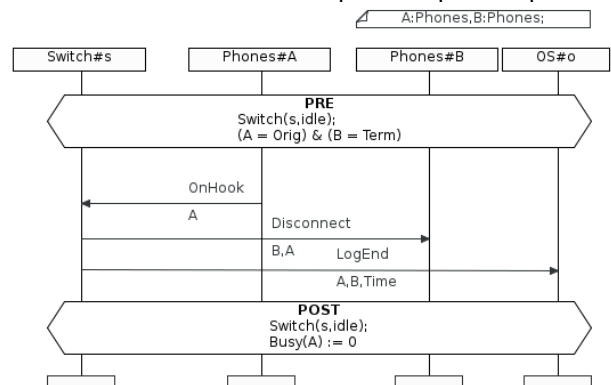


Fig. 1. Example of basic protocol as a part of Plain Old Telephone System

their synchronization. An example of a UCM diagram is given below (Fig 2.).

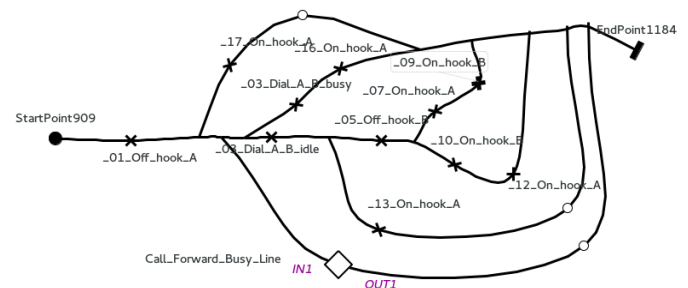


Fig. 2. Use Case Maps for part of telephony protocol

Models that created as composition of UCM present the models of formal requirements that are the source of test generation. This method could be extended for model of

program where UCM defines a control flow of program and basic protocols contains pre- and postconditions for program statements. Such model could be verified with applying of methods of symbolic modelling.

IV. SYMBOLIC MODELING SEMANTICS

The model of a system is considered as a hierarchical set of interacting environments and agents *inserted* (exist) into these environments [13]. The model description has three main levels: basic protocols level, UCM level, and modelling level. Each level can be considered as a level of abstraction in system description. The most abstract is a basic protocols level. It contains the largest behavior describing the evolution of a system. The control level restricts the behavior or set of scenarios strengthening the environment control of agent behaviors according to the general requirements to the system. The modelling level or level of trace generation is the least abstract and depends on a problem solved on a base of the model of a system. It provides further restriction of the set of possible traces generating.

Basic protocol level constitutes with the set of basic protocols. Each basic protocol describes one of the possible elementary scenarios of system behavior. Algebraic form of basic protocol is a formula

$$\forall x (a(x) \rightarrow \langle P(x) \rangle \beta(x)) \quad (1)$$

In this formula $x=(x_1, x_2, \dots)$ is a list of typed parameters, $a(x)$ and $\beta(x)$ are the formulas of logic language called the *basic language* of a model, $P(x)$ is a finite process, which describes the interaction of agents and their environment by message passing. Formula $a(x)$ is called a precondition, and formula $\beta(x)$ is called a postcondition of basic protocol. Basic protocol can be considered as a temporal logic formula that expresses the fact that if (for suitable values of parameters) a system state satisfies the condition $a(x)$, the process $P(x)$ can be activated and, after its successful termination, the new state of a system will satisfy the condition $\beta(x)$. We use MSC language [14] for graphical representation of basic protocols.

Basic language of a model is first order multityped (multisorted) logical language. Simple types are numeric (real and integer), Boolean, symbolic, and enumerated types. There are two kinds of functional symbols in basic language. The first kind consists of the symbols with fixed interpretation (arithmetic operations for numeric, logical connectives for Booleans etc.).

The second kind of functional symbols change their interpretation during the evolution of a system like Abstract State Machine [15]. These functional symbols are called attributes. Attributes of arity zero have simple types and are called simple attributes. The attributes of arity more than zero are called functional attributes. They have fixed types for domain and range values and are used for the representation of data structures like arrays. The access to the values of these data structures provided by the attribute expressions like $f(t_1, t_2, \dots)$ where f is an attribute symbol and t_1, t_2, \dots are attribute or constant expressions. If all arguments of attribute expression

are constant expressions then attribute expression is called a constant attribute expression.

In the set of functional symbols of basic language there are also some distinguished collections of attributes which are called agent attributes. Each collection of agent attributes defines agent type. Agent types define the types of agents, which can be inserted into given environment. The description of environment attributes and the types of agents constitutes environment description and defines the environment type. A pair $\langle E, L \rangle$ consisting of environment description E and the set L of basic protocols consistent with this environment description constitutes the description of a system on basic protocol level.

Let $\langle E, L \rangle$ be the description of a system. For this description we define a transition system $S=S(E, L)$, which describes the evolution and all possible traces of this system.

The state $s=\sigma[u_1, u_2, \dots]$ of a system S is represented as a composition of a state of environment σ and the states u_1, u_2, \dots of agents inserted into this environment. The state of environment can be concrete or symbolic. Concrete state is the mapping from the set of constant attribute expressions (including agent attribute expressions) to the set of their values (consistent with the types of attribute expressions). Symbolic state is a formula of basic language.

The state of agent is defined by the values of constant attribute expressions of agent attributes in concrete state, and the properties of agent attributes in the formula of symbolic state.

Each of basic protocol B defines some transition relation $s \xrightarrow{B} s'$ on the set of concrete system states. To compute this relation one must first instantiate the basic protocol by substituting some concrete values of parameters to basic protocol. If the precondition of basic protocol is true then basic protocol B is applicable to the state s and the state s' is selected non-deterministically among those states on which the postcondition is true. Transition system defined in this way is called a concrete model of local description level. To make a concrete model more deterministic the assignments and conditional assignments are allowed in postcondition. To obtain a trace of a concrete model starting from some initial state s_0 one of the possible histories of a system evolution

$$s_0 \xrightarrow{B_1(z_1)} s_1 \xrightarrow{B_2(z_2)} \dots \xrightarrow{B_n(z_n)} s_n \quad (2)$$

for basic protocols instantiated by z_1, z_2, \dots is used to create a trace $P_1(z_1) * P_2(z_2) * \dots * P_n(z_n)$ as a weak sequential product of instantiated MSC charts for basic protocols processes.

Symbolic model of basic protocols level is a transition system with formulas of basic language as states. Transition $s \xrightarrow{B} s'$ for basic protocol $B = \forall x (a(x) \rightarrow \langle P(x) \rangle \beta(x))$ is computed as follows. First check the satisfiability of the formula $s \wedge a(x)$ where the elements of a list x are considered as the new simple attributes of environment description. If the formula is satisfiable, then compute $pt(s \wedge a(x), \beta(x))$. The function pt is called predicate transformer and it computes the

strongest postcondition provided that precondition before computation was $s \wedge \alpha(x)$. The details for computing the predicate transformer for formulas with quantifiers can be found in [16]. Generating traces from histories is made in a similar way as for concrete model but instantiation is needed only for the names of agents, which are used to name the instances in MSC diagrams.

It is not sufficient to use only basic protocols level for complete specification of a model. The matter is that constraints on sequences of application of basic protocols are not defined on them, which can lead to the consideration of undesirable histories and traces. The next level of a model description consists of the definition of a succession relation on the set of basic protocols. This relation can be introduced by the definition of additional control attributes and conditions limiting the conditions of application of basic protocols on these attributes. An inconvenience of this description is the need for the partition of the basic attributes and auxiliary control attributes. Moreover, the basic protocols themselves become more complicated. Therefore, it is useful to construct the control level as the separate upper level system.

Semantically control system over local description level can be defined as a transition system with the set of actions that includes besides of the own actions the references to basic protocols of a low level system. The control system is considered as an environment for low level system and its state is described by expression $U[s]$ where U is a state of a control system and s is a state of a low level system $S=S(E,L)$. The transition rules for control system over S can be described now by the following rules:

$$\frac{U \xrightarrow{a} U'}{U[s] \xrightarrow{a} U'[s]} \quad a \notin L \quad (3)$$

$$\frac{U \xrightarrow{a} U', s \xrightarrow{a} s'}{U[s] \xrightarrow{a} U'[s']} \quad a \in L \quad (4)$$

If we replace the first rule by the rule

$$\frac{U \xrightarrow{a} U'}{U[s] \rightarrow U'[s]} \quad a \notin L \quad (5)$$

with hidden transition then an external observer will not find any difference between the functioning of the low level system with control and without control. He will see only traces, and they are traces of the same low level system. However, with the use of the control system, their number can be smaller, and also deadlock states can arise that are absent in the case of the low level system if the control system is not correct.

V. SYMBOLIC VERIFICATION

We consider the verification of requirements for reactive systems presented as combination of UCM notation and basic protocols.

It supports the following kinds of verification. *Checking the consistency* of requirements means detection of non-determinism and ambiguities in behavioral requirements. Often

such issues are deeply hidden in specifications and could entail subsequent errors and misunderstanding by developers. *Incompleteness detection* helps finding of deadlock situations in formal model of requirements.

Different from concrete modeling, symbolic methods involve deductive systems such as provers or solvers. Formal model could present the system at a high level of abstraction where deductive techniques are most suitable. Deductive systems provide proofs of assertions in a first-order theory, resultant from the integration of theories of integer and real linear inequalities, enumerated data types, uninterpreted function symbols, and queues. This technique allows the verification of the model for systems with a large or infinite number of states.

Symbolic techniques also support incremental verification that is important for the development of large systems. Different parts of formal specifications can be verified separately and encapsulated into enlarged entities to avoid repeating the verification of such components. For a system with high degree of features interactions each feature can be verified separately first and their interactions can be verified without examining the individual behaviors repeatedly.

Author is one of the developers of IMS (Insertion Modeling System) developed in Glushkov Institute of cybernetics. Symbolic simulation in IMS allows detecting reachability of the violation of correctness properties by considering symbolic states of a system and generating a set of traces leading to the findings.

A static requirements checking in IMS is intended for detection of candidates for properties violations. It is based on formal proving of statements and involves deductive tools. If property violation is detected then its reachability shall be proved by means of symbolic modelling.

The system establishes a trace that leads to a deadlock or other anomaly situation and identifies its causes. A trace is graphically presented as a MSC diagram. Safety violations in IMS also are detected by means of symbolic modeling or static proving. Liveness of a system is checked by finding the reachability of the necessary property. Livelock detection identifies situations where a system may or otherwise be non-responsive.

VI. SYMBOLIC TEST GENERATION

The formal presentation of requirements in UCM notation together with basic protocols gives the possibility to generate a test suite at the given level of abstraction in IMS system. A set of traces can be generated from formal requirements that will be obtained in MSC notation and can be converted into standard test formats.

Traces contain input and output signals and local actions of the system together with the set of states of the environment that contains possible values for the system attributes. These states are symbolic and cover potentially large sets of attribute values. The generation of traces corresponds to different coverage types defined in a *user trace generation request*.

Node coverage. Coverage of all nodes constructs in UCM diagram that corresponds to coverage of all functionalities of model or coverage of requirements.

Edge coverage. Coverage of all adjacent pairs of UCM responsibility nodes or other transition points.

Path coverage. Coverage of all paths those are possible between start and end points with restriction on the length of a path or the number of visiting the same nodes of UCM. Path is a sequence of “responsibility” constructs presented by basic protocols.

Full state coverage. Coverage of all states of the system that can be reached by the restricted number of steps.

The request for trace generation could be extended by selection of starting or end points or excluding of UCM nodes or edges. The traversal of constructs “stubs” that present references to other UCM diagrams also could be tuned by definition of mentioned coverage type for itself. The strategy also could be defined by length of generated traces. It could be generated as the shortest distance between start and end points or cover the maximal number of applied basic protocols.

Unreachable edges of UCM diagrams could occur during trace generation. It could happen due to the insufficient adjustments of trace generation for instance insufficient number of loops or maximal length of trace.

After trace generation, there are number of unreachable nodes and edges of UCM diagrams. Their unreachability should be proved or refuted by the following possibilities:

Backward trace generation. We generate backward trace by means of backward predicate transformer from candidates for unreachability to known reachable points. If backward generation will finished at finite number of steps by traversal of all possible paths and will not reach given point, then this point is unreachable.

Invariants usage. During trace generation the set of invariants could be generated for every basic protocol. We consider preinvariant – formula that defines possible states of environment before basic protocol application, and corresponding postinvariant after. We consider invariant as overapproximation of set of states or the weakest invariant formula. If intersection of invariant formula and precondition of successor is empty then the successor is unreachable.

All generated traces forms the test suite for further usage in test execution. Every trace is MSC diagram marked by the formulas of environment state.

Test could be built by selection of the instance for SUT (system under testing) and instance for testing system in MSC diagram. Message in MSC diagram defines the points of interaction between SUT and testing system. Formula defines the set of possible parameters of output message from testing system and oracle set of values for comparison with input messages as reaction of SUT.

For testing procedure the concretization of attributes shall be provided. It is performed by solving of constraints presented

by given environment formulas. Then testing procedure or test execution is implemented by launching of tested program under control of input values from generated tests. Test suite could be converted to standard test format for instance TTCN format and be executed on existing standard test execution tools.

VII. SYMBOLIC TEST EXECUTION

Method of concretization of symbolic traces is useful when test model and tested program are on the same level of abstraction. If test model is given on a higher level of abstraction then concrete values could not be defined correctly. The other disadvantage is that some scenario of behavior could be missed due to the high degree of detailing of program code.

To overcome these disadvantages the symbolic test execution is proposed. It based on the symbolic execution of tested program correspondingly to generated test.

The scheme of symbolic testing is the following.

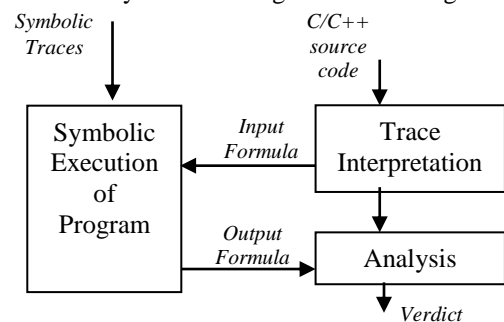


Fig. 3. Scheme of Symbolic test execution

The input of symbolic testing procedure is the set of generated symbolic traces as MSC marked by formulas over parameters of messages, and source code of system to be tested. The program of symbolic execution starts SUT performing correspondingly to scenario given in MSC trace. It changes the code environment under control from tested system that is also defined by formula over program variables. SUT accepts input data as formulas and symbolically executes statements. It returns the output as formula of code environment state and the following analysis is performed.

Let us consider some cases of symbolic test execution. Let the oracle of the code environment is a formula X and symbolic state of code environment is Y .

- If the conjunction of Y and $\neg X$ is not satisfiable then test is passed.
- If the conjunction of Y and $\neg X$ is satisfiable then there exist an unpredictable concrete state not covered by X and the formula $Z=Y \vee \neg X$ covers all such states.

In the last case, there are two possibilities.

- The conjunction of X and Y is unsatisfiable (equal to 0). All possible concrete states are unpredictable, the testing process failed.

- The conjunction of X and Y is satisfiable. The testing process can be continue, however system signalizes about the existence of unpredictable states.

Backward symbolic moving via given test from the current set of unpredictable states of environment covered by the formula Z allows detecting the error cause.

Symbolic test execution is based on symbolic execution of program under control of test. We consider C/C++ source code and encode it to the UCM notation with basic protocols. The corresponding interpretation in the term of basic protocols is implemented for C/C++ control statements, assignment and condition statements with linear arithmetic, calls of functions, and all kinds of cycle statements. Address arithmetic and pointers, structures, arrays, unions with integer, long, real, character, double data types also are realized. The pre- and postconditions were created for functions from input/output, memory, standard libraries.

We consider two next kinds of symbolic test execution. The first is symbolic “black box” online testing where we use the test model for generation of behavior of SUT and compare its generated oracles with symbolic environment of the executed program code. In this case we are trying to generate coverage for test model disregarding code coverage.

The scheme of “black box” online testing is the following.

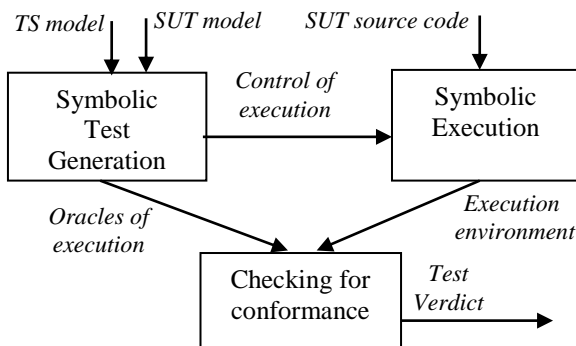


Fig. 4. “Black box” online symbolic testing

Here we generate the traces for parallel composition of TS (tested system) and SUT (system under testing). The generation is controlled by symbolic execution of SUT source code correspondingly to generated traces. The obtained states of environment are compared for conformance in checking module.

There are a number of requirements for implementation of such technique. One of this is that the level of abstraction for interchanged signals for source code shall be the same as for test (requirements) model. It means that the set of signals shall be the same for both models. The correspondence between names of variables of code and names of system attributes also shall be defined.

The other kind of online testing is symbolic “white box” testing where we execute source code symbolically checking the conformance of constraints from generation in

requirements model and symbolic environment of executed program source code. In this case we are trying to cover all branches of source code. The scheme of “white box” testing is the following:

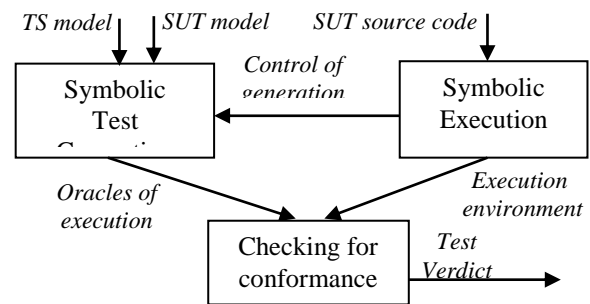


Fig. 5. “White box” symbolic testing

The difference is that source code symbolic execution controls the trace generation and compare for conformance the corresponding constraints.

Note that for requirements models of high level of abstraction the symbolic test execution is insufficient. In this case test execution with concrete values could be useful additionally.

REFERENCES

- [1] J. C. King, "A new approach to program testing," in Proc. Int. Conf. Reliable, Software, Apr. 1975, pp. 228-233.
- [2] Doron Peled, Patrizio Pellicone, Paola Spoletini, "Model Checking", Wiley Encyclopedia of Computer Science and Engineering, 2009.
- [3] Kenneth L. McMillan, "Symbolic Model Checking", Kluwer Academic Publisher, 1993.
- [4] Patrick Cousot, "Formal Verification by Abstract Interpretation", Lecture Notes in Computer Science, 2012, vol. 7211, pp. 3-7, Springer.
- [5] Robert B. Jones, "Symbolic Simulation Methods for Industrial Formal Verification", 2002, Springer.
- [6] Edmund Clarke, Orna Grumberg, Somesh Jha, Yuan Lu, Helmut Veith, "Counterexample-Guided Abstraction Refinement", Lecture Notes in Computer science Volume, 1855, 2000, pp. 154-169.
- [7] J.Peleshka, E.Vorobev, F.Lapschies, C.Zahlten, "Automated Model-Based Testing with RT-Tester", Technical report, 25.05.2011.
- [8] C. Pasareanu, N.Rungta, "Symbolic PathFinder: symbolic execution of Java bytecode", Proceeding ASE '10, pp.179-180.
- [9] Patrice Godefroid, Michael Y. Levin, David Molnar, "SAGE: Whitebox Fuzzing for Security Testing", Magazine Queue-Networks, Vol. 10, Issue 1, 2012.
- [10] Guodong Li, Indradeep Ghosh, Sreeranga P. Rajan KLOVER: "A Symbolic Execution and Automatic Test Generation Tool for C++ Programs", Lecture Notes in Computer Science, Volume 6806, 2011, pp.609-615.
- [11] ITU-T Recommendation Z.151, "User Requirements Notation (URN) – Language definition", 10.2012.
- [12] A.A. Letichevsky, J.V. Kapitonova, V.A. Volkov, A.A. Letichevsky jr., S.N. Baranov, V.P. Kotlyarov, T. Weigert "System Specification with Basic Protocols", Cybernetics and System Analyses. 2005, № 4, pp. 3–21.

- [13] Letichevsky A., Gilbert D., “A Model for Interaction of Agents and Environments”, Lecture Notes in Computer Science, 1999, №182, pp. 311-328.
- [14] ITU-T Recommendation, Z.120, Message Sequence Charts.
- [15] E. Borger, R. Stark, ”Abstract State Machines”, Springer-Verlag, 2003.
- [16] A. Letichevsky, A. Godlevsky, O. Letychevskiy, S. Potienko, V.Peschanenko, “The properties of predicate transformer of the VRS system”, Cybernetics and System Analyses, №4, 2010, pp.3-16.

Design, Implementation and Comparison of Low-Cost Laser Scanning Systems for 3D Modeling

Tuba Kurban

Dept. of Geomatics Engineering
Erciyes University
Kayseri, Turkey
tubac@erciyes.edu.tr

Erkan Besdok

Dept. of Geomatics Engineering
Erciyes University
Kayseri, Turkey
ebesdok@erciyes.edu.tr

Abstract—3 dimensional (3D) modeling of an object or an environment using point clouds is an important problem in many scientific fields such as photogrammetry, remote sensing, materials processing, reverse engineering, construction industry, virtual reality and medicine etc. Laser scanning is an effective technique that facilitates 3D modeling process with providing large amount of 3D point cloud data in a short time. In this study, design process of point laser sensor and line laser sensor based low cost scanner systems is proposed. Performed 3D data measurements with these two different laser scanners show that; point laser range sensor based scanner, that can capture lesser 3D point for per second, provides more detailed and more sensitive measurements. It can be preferred in applications when the details are very important and are suitable for modeling small objects. However, line laser range sensor based scanner can capture much more 3D point data per second and it is suitable for applications where time critical models with large objects and environment.

Keywords—laser scanning; 3d data acquisition; point cloud; 3d modeling

techniques, laser scanners provide large amount of 3D data more precisely and quickly [20].

I. INTRODUCTION

Laser scanners can obtain large amounts of data called 3D point cloud and this data is used for constructing a 3D model which is used to facilitate the analysis of a real world object or an environment [1, 2]. 3D scanning technology is used variety of fields such as photogrammetry and remote sensing [3], materials processing and manufacturing [4-8], reverse engineering [5-8], civil engineering [9], virtual reality and augmented reality [10], cultural heritage [11, 12] and medical [13] applications.

Wide variety of hardware and software based solutions for 3D data acquisition is commercially available. Laser scanner [14], structured light scanner [15], stereo vision [16], photogrammetry [17], interferometry [18] and shape from shading [19] are popular techniques have been developed. With the advances in technology, accuracy of these devices is increased while costs are decreased. This situation makes 3D data attractive for many applications. Among all these

In this study, two different scanner systems are designed that differs in terms of used laser range sensor type. Point laser and line laser range sensors are mounted on a highly precise pan tilt unit. Coordination, between pan tilt unit that performed the localization and laser range sensors, is controlled precisely by developed 3D data acquisition software.

At the rest of this paper proposed point laser based scanner and line laser based scanner are discussed in section 2. Comparison of the two different laser range sensor and measurement results are given in section 3. Conclusions are in section 4.

II. PROPOSED LASER SCANNING MODELS

Proposed low cost scanning systems consist of a pan tilt unit and two different laser range sensors. Pan tilt unit that used in our system is PTU D46-70 [21] obtained from Directed Perception. PTU-D46-70 can be controlled from a

computer via RS-232 port. The operation speed of this unit is 60° / sec and resolution is 0.012857° per pan and tilt step. Load capacity is over 4.08 kg. Therefore, it is preferred in many applications such as robotics, computer vision, security, surveillance, industrial automation, tracking, webcams and laser ranging. PTU-D46-70 pan tilt unit is given in Fig. 1.

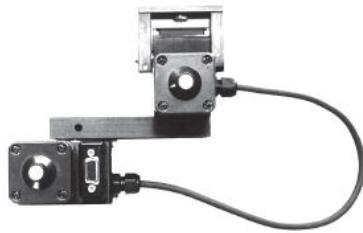


Fig. 1. PTU D46-70 pan tilt unit.

Our measurement systems differ in terms of laser scanner sensor types. First laser range sensor that used in our system is FLS C-10 [22] point-laser scanner obtained from Dimetix. Point laser range sensor only detects the distance of one point in its direction of view. The other sensor that used in the system is UTM 30 - LX [23] obtained from Hokuyo. The sensor view direction can be changed by using a rotating mirror. Thus, scanning process performed along a line.

In our measurement system, the pan-tilt system has been placed on a tripod. Each laser range sensor can be mounted on the pan tilt unit. Experimental setup can be seen in Fig. 2. Laser sensors measures continuously while pan tilt unit take steps along maximum and minimum pan and tilt range of motion. PTU-D46-70 pan-tilt has 24836 steps for pan and 5823 steps for tilt. In other words, it can scan -159.65° to 159.67° for pan and -45.1° to 29.76° for tilt. Both pan tilt unit and laser range finder are connected to a host computer and can be controlled from RS-232 port. The details of the proposed models are given in the following sub-sections:

A. Point-laser based scanner

FLS C-10 is a powerful laser range sensor for industrial applications. It allows accurate and contactless distance measurement over a wide range using the reflection of a laser beam. It measures from 0.05 to 65 meters with 1.0 millimeter accuracy. The laser distance range finder is a safe class II laser device that can measure both near and far. Technical specifications of the sensor are given in Table 1 and FLS C-10 point laser range sensor is given in Fig. 2(a).

TABLE I. TECHNICAL SPECIFICATIONS OF FLS C-10POINT LASER.

Technical Specifications	
Measuring Range	0.05 m - 65 m
Accuracy	± 1.0 mm
Max. Measuring Rate	200Hz
Dimensions	150 x 80 x 55 mm
Weight	690g
Serial interfaces	RS-232, RS-422
Signal Measurement	Single or Continuous

Dimetix FLS-C10 sensor is working with phase shift principle. The sensor scans the scene with a periodic light wave and distance to the scene is calculated from the difference between the phase and the frequency of the received signal. Because of no waiting for the received signal, this method provides high sample density and accuracy in comparison with time of flight principle.

B. Line-laser based scanner

Hokuyo UTM-30LX is a Class I laser distance measurement sensor for many applications that require high speed and accuracy. The sensor can be changed view direction by a rotating mirror and thus measure the points along a line at a time. Hokuyo UTM-30 LX is working with the time of flight principle. This principle is based on the calculation of the time between the transmitted and received laser signals as a result of surface reflection. The distance of the desired point is calculated from speed of light and round trip delay of the signal. UTM-30LX line laser sensor has a rotating mirror and position encoder. With the rotating mirror scanning process is performed along a line and position encoder determines the measurement resolution on the line. The standard encoder resolution is 1080 counts per revolution.

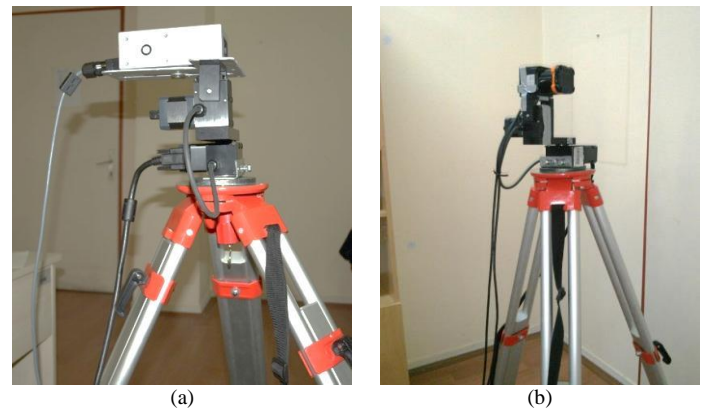


Fig. 2. Proposed laser scanner systems: (a) point laser, (b) line laser.

TABLE II. HOKUYO UTM-30LX TABLE 1. TECHNICAL SPECIFICATIONS OF HOKUYO UTM-30LX

Technical Specifications	
Measuring Range	0.1 ~ 30m
Accuracy	0.1 - 10m:±30mm, 10 - 30m:±50mm
Max. Measuring Rate	40 Hz.
View	270°
Dimensions	60 x 60 x 87mm
Weight	210 gr
Serial interfaces	RS-232, RS-422
Signal Measurement	Single or Continuous

Line laser range finder is mounted up to the pan tilt unit in our system as shown in Fig. 2(b). However, tilt motion is performed by the rotating mirror. Only pan motion can be controlled by the used in this type of scanner.

C. 3D Data Acquisition System

Both in point and line laser based scanners, the output is pan angle, tilt angle and distance. The analytical 3D model is given in Fig. 3. In Fig. 3, origin of the coordinate system is intersect of the two axes of control platform X and Y. Z axis is parallel to laser distance measuring line. M(x,y,z) is the measured point and D is the distance between point and control platform. t (tilt angle) is the vertical scanning step angle of control platform and the p (pan angle) horizontal scanning step angle of control platform. 3D coordinates of the measured point can be calculated from pan angle, tilt angle and distance by Equation 1.

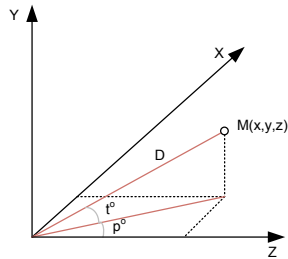


Fig. 3. Analytical 3D model

$$\begin{aligned} y &= D \cdot \sin(t) \\ x &= D \cdot \cos(t) \cdot \sin(p) \\ z &= D \cdot \cos(t) \cdot \cos(p) \end{aligned} \quad (1)$$

Pan-tilt unit and laser sensors can be controlled from serial port. Developed 3D data acquisition system, provide the control of the pan-tilt unit and laser sensors, according to the user specified parameters. Data acquisition system contains a variety of programs that developed in MATLAB environment to receive, record and process of the pan angle, tilt angle and distance data. In 3D data acquisition systems, user can control scanner system with the parameters of scanning range, step size for pan and tilt, start and finish positions of the pan and tilt. Operation steps of the system can be seen in Fig. 4.

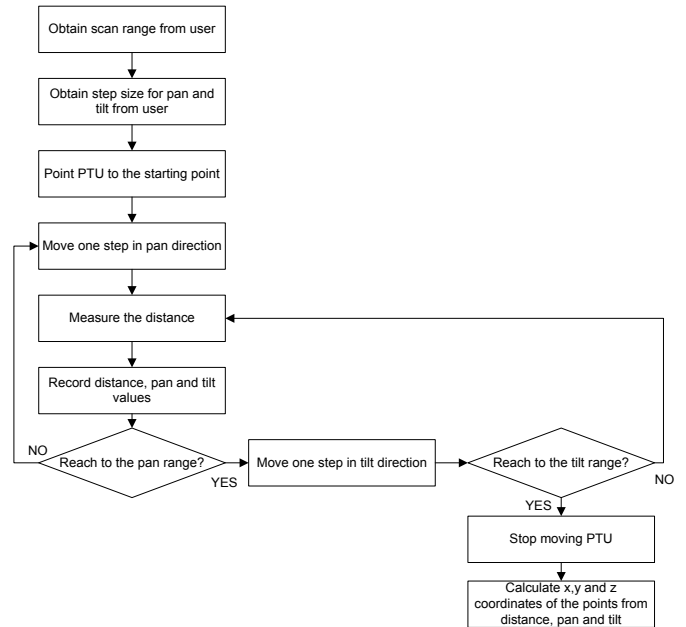


Fig. 4. Flowchart of laser scanner 3D data acquisition software

III. 3D MEASUREMENTS WITH PROPOSED LASER SCANNERS

Two different laser sensors are used in our system. Point laser can measure a single point at the view of the direction. Localization in horizontal and vertical direction is performed by pan tilt unit. On the other hand, line laser can measure along a line at the view of the direction. Vertical position derived from rotating mirror and position encoder in its structure. Horizontal localization is performed by pan tilt unit. In point laser scanner both in horizontal and vertical positioning depends on the pan tilt unit resolution.

Advantages and disadvantages of both scanner systems are listed below:

- Point laser scanner can measure a single point; on the other hand, line laser scanner can measure the points along a line at a time.
- Point laser scanner is more precise because of the positioning procedure with pan tilt unit both in horizontal and vertical. Line laser sensor has a same resolution in horizontal direction; however the positioning resolution depends on encoder resolution.
- Point laser scanner can capture 20 point in a second. Pan tilt unit resolution is 0,012858° and this unit can get 24836 different positions for pan and 5823 different position for tilt.
- Line laser has a rotating mirror and 1080 count per revolution. With 25 Hz measuring rate, it can capture 27000 point at a time. Only pan motion can be

controlled by user. Pan angle can get 24836 different values.

- Point laser scanner is suitable for applications that details are very important. Scanning time is reasonable for small object measurements.
- Line laser scanner is suitable for applications that require less detail. Scanning time is reasonable for large objects or environments. In particular, applications such as cultural heritage, environment reproduction, indoor and outdoor mapping where the scanning area is large.

For the assessment of scanning performances a small *sculpture* is modeled with point laser scanner and a *laboratory* is modeled with line laser scanner. 3D point clouds belong to same object or environment taken from different locations laser scanners are registered with well-known Iterative Closest Point algorithm. After registration, obtained noisy point cloud that represents the whole object, is filtered by singular value decomposition method. Obtained point clouds are given in Fig. 5 and 6 for *sculpture* and *laboratory* model, respectively.

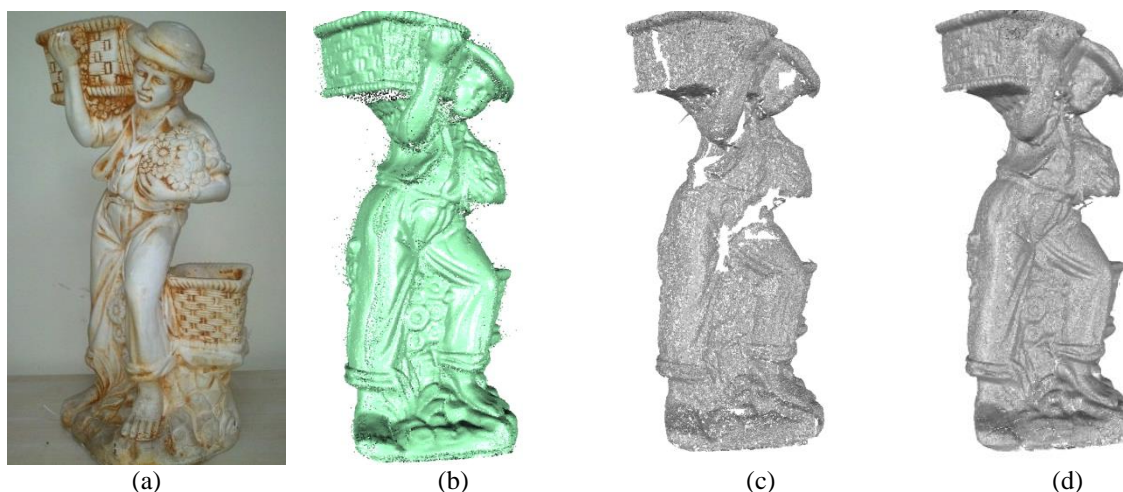


Fig. 5. Sculpture model: (a) original model, (b) 3D point cloud, (c) noisy mesh, (d) filtered mesh.

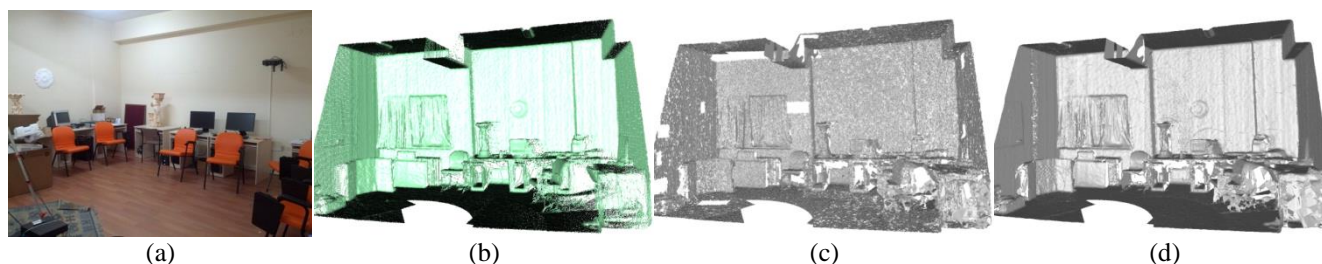


Fig. 6. Laboratory model: (a) original model, (b) 3D point cloud, (c) noisy mesh, (d) filtered mesh.

IV. CONCLUSION

In this study, two different laser scanning systems are proposed. Scanner systems are differing in terms of laser range finder type. Measurement results showed that point laser based scanner is slower than line laser based scanner. Point laser range finder scanned the sculpture very detailed and scanning time is acceptable. The sculpture is a small object and scanning area is small. Point laser based scanner provides more detailed and more sensitive for scanning small objects and in applications when the details is very important.

Scanning of the laboratory is very easy with line laser and scanning time is very short. This type of sensors can be preferred applications such as cultural heritage, built environment, indoor mapping etc.

ACKNOWLEDGEMENT

The studies in this paper have been supported by Erciyes University FBA-10-3067, FBA-9-1131.

REFERENCES

[1] T. Al-Hawari, *et al.*, "2D laser scanner selection using fuzzy logic," *Expert Systems with Applications*, vol. 38. 2011, pp. 5614-5619.

- [2] Z. Xie, S. Xu, and X. Li, "A high-accuracy method for fine registration of overlapping point clouds," *Image and Vision Computing*, vol. 28. 2010, pp. 563-570.
- [3] Y. Arayici, "An approach for real world data modelling with the 3D terrestrial laser scanner for built environment," *Automation in Construction*, vol. 16. 2007, pp. 816-829.
- [4] M. Mahmud, *et al.*, "3D part inspection path planning of a laser scanner with control on the uncertainty," *Computer-Aided Design*, vol. 43. 2011, pp. 345-355.
- [5] S. Son, H. Park, and K. H. Lee, "Automated laser scanning system for reverse engineering and inspection," *International Journal of Machine Tools and Manufacture*, vol. 42. 2002, pp. 889-897.
- [6] M. Korosec, J. Duhovnik, and N. Vukasinovic, "Identification and optimization of key process parameters in noncontact laser scanning for reverse engineering," *Computer-Aided Design*, vol. 42. 2010, pp. 744-748.
- [7] S. Larsson and J. A. P. Kjellander, "Path planning for laser scanning with an industrial robot," *Robotics and Autonomous Systems*, vol. 56. 2008, pp. 615-624.
- [8] S. Larsson and J. A. P. Kjellander, "Motion control and data capturing for laser scanning with an industrial robot," *Robotics and Autonomous Systems*, vol. 54. 2006, pp. 453-460.
- [9] G. L. Heritage and D. J. Milan, "Terrestrial laser scanning of grain roughness in a gravel-bed river," *Geomorphology*, vol. 113. 2009, pp. 4-11.
- [10] C. Portalés, J. L. Lerma, and S. Navarro, "Augmented reality and photogrammetry: A synergy to visualize physical and virtual city environments," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 65. 2010, pp. 134-142.
- [11] D. Miyazaki, *et al.*, "The great buddha project: Modeling cultural heritage through observation," in *Modeling from Reality*. vol. 640, K. Ikeuchi and Y. Sato, Eds., ed: Springer US, 2001, pp. 181-193.
- [12] N. Yastikli, "Documentation of cultural heritage using digital photogrammetry and laser scanning," *Journal of Cultural Heritage*, vol. 8. 2007, pp. 423-427.
- [13] K. Schwenzer-Zimmerer, *et al.*, "Quantitative 3D soft tissue analysis of symmetry prior to and after unilateral cleft lip repair compared with non-cleft persons (performed in Cambodia)," *Journal of Cranio-Maxillofacial Surgery*, vol. 36. 2008, pp. 431-438.
- [14] G. Sithole and G. Vosselman, "Experimental comparison of filter algorithms for bare-Earth extraction from airborne laser scanning point clouds," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 59. 2004, pp. 85-101.
- [15] S. Izadi, *et al.*, "KinectFusion: real-time dynamic 3D surface reconstruction and interaction," presented at the ACM SIGGRAPH 2011 Talks, Vancouver, British Columbia, Canada, 2011.
- [16] D. Samaras, *et al.*, "Variable albedo surface reconstruction from stereo and shape from shading," presented at the IEEE Conference on Computer Vision and Pattern Recognition, South Carolina, USA, 2000.
- [17] K. A. Hashim, *et al.*, "Integration of low altitude aerial & terrestrial photogrammetry data in 3D heritage building modeling," presented at the Control and System Graduate Research Colloquium (ICSGRC), 2012 IEEE, 2012.
- [18] P. Kasprzak and P. Kowalczyk, "Objects recognition with high-resolution in SAR data and global geometric feature map," in *Signal Processing Symposium (SPS), 2013*, 2013, pp. 1-5.
- [19] J. J. Atick, P. A. Griffin, and A. N. Redlich, "Statistical approach to shape from shading: reconstruction of three-dimensional face surfaces from single two-dimensional images," *Neural Computation*, vol. 8. 1996/08/15 1996, pp. 1321-1340.
- [20] G. Pavlidis, *et al.*, "Methods for 3D digitization of cultural heritage," *Journal of Cultural Heritage*, vol. 8. 2007, pp. 93-98.
- [21] FLIR. (2014, 01.06.2014). *PTU D46-70 pan-tilt*. Available: <http://www.flir.com/mcs/view/?id=53712>
- [22] Dimetix. (2014, 01.06.2014). *FLS-C10 Lazer*. Available: http://www.dimetix-usa.com/catalog/product_info.php?cPath=22_43&products_id=88
- [23] Hokuyo. (2014, 01.06.2014). *UTM-30LX Lazer*. Available: https://www.hokuyo-aut.jp/02sensor/07scanner/utm_30lx.html

Plane Segmentation of Kinect Point Clouds using RANSAC

Rifat Kurban

Dept. of Computer Engineering
Erciyes University
Kayseri, Turkey
rkurban@erciyes.edu.tr

Florenc Skuka

Dept. of Computer Engineering
Erciyes University
Kayseri, Turkey
skuka.f@gmail.com

Hakki Bozpolat

Informatics and Inf. Security Research Center
TUBITAK
Gebze, Turkey
hakki.bozpolat@tubitak.gov.tr

Abstract—In this paper segmentation of planes in point cloud data generated by Microsoft Kinect is detected by using RANSAC method. Two experimental data are acquired by OpenNI and OpenCV library. Kinect camera is first calibrated and the holes in the acquired data are filled. Then, the data is filtered, downsampled and segmented via Point Cloud Library (PCL). Distance threshold and normal weighted distance parameters of RANSAC algorithm are evaluated in the experiments.

Keywords—plane segmentation, Kinect, point cloud library

I. INTRODUCTION

In the last few years three dimensional point cloud processing has become an exciting and very hot research topic in computer vision. Compared to 2D data, range images have several advantages such as 3D geometry processing, 3D shape matching and modeling, feature extraction and matching, segmentation, and object recognition [1]. All these advantages in point cloud data processing are capturing the attention of developers and researchers to develop 3D based applications [4] in many computer vision fields, such as: extracting the user silhouette [4] people detecting and tracking [2], 3D human body modeling and shape analyzing, especially for virtual shopping and clothing industry [3], 3D sensing of the environment representations [5] etc.

Recent years many mobile robots have been developed to help people in their works. In order to help humans in their life, first of all, a robot has to percept environments as it is in 3D after that it can process the data according to the job that it has to do. In 3D model environments flat surfaces are really common, as well as to their attractive geometric properties, therefore plane segmentation in such environments is an essential task. In computer vision one of the common algorithms for detecting planes is Random Sample Consensus (RANSAC) [6] which consists to search for the best plane among the point cloud data. Providing depth data with low-cost is made able by last developed sensors such as Microsoft Kinect and Asus Xtion PRO.

In [9] a plane detection method in point cloud data is proposed, which detect planes by integrating RANSAC method and minimum description length (MDL). This method can avoid detecting wrong planes in point clouds with complex geometry. It follows these steps to detect planes: It divides point cloud in small rectangles block, in each block applies RANSAC for detecting all planes, and then MDL is used to decide how many planes are in each divided block.

Different versions of the Hough Transform [11] to detect planes in 3D point cloud data are evaluated [10]. An accumulator design to achieve the same size per each cell is presented. This method gives good results when Randomized Hough Transform is applied on 3D data. Because, removing all points which lie on the detected plane, increase the performance for detecting next plan. Moreover, the outcomes show that detecting the underlying structures on the search space is preferred as it can be useful to detect large planes.

In [12] a real-time plane detection method based on depth map from Microsoft Kinect sensor is proposed. A system to detect multiple planes fast and roughly in point cloud data acquired from Kinect is suggested. In order to achieve good results for fast detecting multiple planes in 3D point cloud data, they compute local normal vectors of whole point cloud data and classify points in different planes using these local normal vectors. This method has execution time of 2ms and error of 1~2mm levels also shows that it is faster than 3D Hough Transform and RANSAC for plane detection and works in real time.

A plane detection method for image sequences acquired from Kinect sensor is presented in [13]. Image sequences rather than a single image like other methods, which provide higher accuracy and robustness results, are used. This method considers the limitation of the depth sensor by using visual data to help detecting planes. Dealing with sequence of images helps accuracy and gives odometer information.

II. DATA ACQUISITION

Data acquisition includes both hardware and software systems. There are many types of hardware (depth sensors) such as: stereo vision camera, 3D time of light sensor structure light based camera [8], which can capture depth data in different environments. Software system is needed for processing the data to be functionally meaningful information. Transformation of the data to the required model for an

application needs to pass through some steps: data filtering, data registration and integration, surface reconstruction, data simplification and smoothing, feature detection, data segmentation and data compression [7]. Several techniques are used to acquire depth images. By using just RGB camera it is able to reconstruct 3D environment, however this method requires significant amount of post-processing [8]. Another way for acquiring 3D data is using laser scanning, but this device's cost is high. Recently low-cost depth sensors such as Microsoft Kinect and Asus Xtion have become widely available and compared to stereo cameras the quality of depth data has been improved.

Microsoft Kinect sensor can acquire depth data in lightless environments. Kinect camera was first release by Microsoft on November, 2010 as it can be seen in Fig.1. Microsoft Kinect sensor contains two cameras and one laser-based IR projector. It is able to produce 640x480 pixels 32-bit color images and 320x240 pixels 16-bit depth images both at 30 frames per second. Depth images are provided by projecting light patterns on the surrounding scene, IR camera receives the reflected light and compare their positions with the reference pattern [4]. Mostly Kinect has advantages but has some disadvantages too like limitation in detection, Kinect cannot acquire depth image in distance less than 50 cm and more than 10 m, but in order to get good precision we take into consideration data in the range of 0.5 meters up to 8 meters. Also due to the occlusion of IR projection in depth images there are some missing regions, non-measured depth pixel [13]. Another disadvantage of Kinect is that it cannot acquire depth images under sunlight. In order to improve the quality of depth images, acquired data should pass through some filters.

In this work as input source for data acquiring we adopt Microsoft Kinect Sensor with OpenNI (Open Natural Interaction) driver [15] which enables communication with RGB and depth cameras of Kinect, and OpenCV (Open Source Computer Vision) library [16] which is used for processing depth images.

OpenNI is an open source multi-languge, cross-platform framework which provides an application programming interface for writing applications utilizing natural interaction. OpenCV is open source library that supports different applications in computer vision. Moreover it is free for both commercial and non-commercial use. This library has different interfaces such as: C, C++, Java and Python, it also supports Windows, Linux, Mac OS, iOS and Android.



Fig. 1. Microsoft Kinect Sensor v1.

III. POINT CLOUD DATA PROCESSING

To achieve robust results from data provided by Microsoft Kinect sensor it is very necessary to apply some operations on these raw data. In Fig 2., the way how to process with point cloud data is shown.

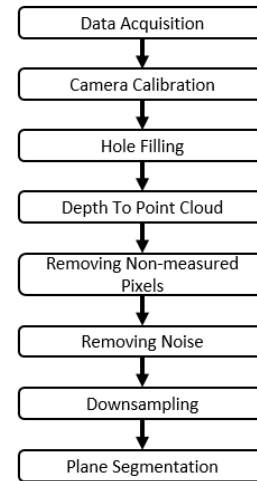


Fig. 2. Flow Chart of Point Cloud Processing For Plane Segmentation.

Camera calibration is a necessary step in computer vision for tasks such as 3D reconstruction. The camera calibration procedure consists of estimating intrinsic parameters, extrinsic parameters and distortion coefficients. Intrinsic parameters consist of camera focal length, and principal point, whereas extrinsic parameters stands for rotation (R) matrix and translation vector (t) of the sensor with respect to the world coordinate system. Distortion coefficients are the coefficients from the radial and tangential distortion.

Hole filling, is needed due to the occlusion of IR projection, acquired depth images have some non-measured depth pixels. Before working with such data it is necessary to fill these holes.

Depth to point cloud is used for transforming a depth images to point cloud data. This can be done by using IR camera intrinsic parameters, focal length and principal point. The algorithm for converting depth images to point cloud data is as following:

Algorithm 1: Depth To Point Cloud

```

1: factor = 1000; //Meter to MM
2: for v=1 to height
3:   for u=1 to width
4:     z = depth(v,u) / factor;
5:     x = (u - cx_d) * z / fx_d;
6:     y = (v - cy_d) * z / fy_d;
7:     pcloud(v,u,1)=x;
8:     pcloud(v,u,2)=y;
9:     pcloud(v,u,3)=z;
10:  end
11: end
    
```

where fx_d , fy_d are the focal length and cx_d , cy_d are the principal points.

Removing Non-measured Pixels is needed because even if some filters are applied for filling holes in point cloud we can see that not always the whole holes are filled. This depends on the size of *nmp* (non measured pixels) in point cloud. In order to prevent these non measured pixels effecting negatively the results, removing these *nmp* pixels is necessary. Furthermore removing *nmp* changes the structure of point cloud from organized to unorganized. In the same time, removing the *nmp* will change the size of the point, and in the unorganized point cloud data the height is set to 1 and width size is the rest of the points.

Removing Noise, decreases unwanted points from raw data provided by Kinect. For robust processing of point cloud data it is important and necessary to apply some filters for removing noise and outliers out of the original point cloud data because they can effect and produce errors in processing. As a result the point cloud data is classified as inlier point and outlier points. After filtering, point cloud data contains just inlier points. Moreover removing outliers from point cloud decrease the processing time as well as downsampling.

Downsampling, also reduces the amount of points. Microsoft Kinect sensor produces a point cloud containing 307200 (640x480) points. Working with such high resolution point clouds for detection planes requires a lot of processing time.

In order to decrease the process time of detection planes in point cloud data applying some optimization techniques such as Voxel Grid Downsampling filter is required. Setting parameters in efficient way to voxel grid filter will yield very good results also will reduce execution time and cost of the CPU power.

Plane segmentation, is the final step. After point cloud data is processed RANSAC based plane fitting method is applied to detect planes in point cloud data robustly. RANSAC method finds the largest set of points that fit to plane. The plane equation in three dimensional point cloud data can be defined as:

$$ax + by + cz + d = 0 \quad (1)$$

Where *a*, *b* and *c* are plane parameters and *d* is distance of plane from the origin. RANSAC selects randomly three points from dataset and calculates the parameters of the corresponding plane, after that tries to enlarge the plane according to a given threshold, [17].

Algorithm 2: Ransac

```

1: bestSupport = 0; bestPlane(3,1) = [0, 0, 0]
2: bestStd = ∞; i = 0
3: ε = 1 - foreseeable-support/length(point-list)
4: N = round(log(1 - α)/log(1 - (1-ε)3))
5: while i ≤ N do
6:   j = pick 3 points randomly among (point-list)
7:   pl = pts2plane(j)
8:   dis = dist2plane(pl, point-list)
9:   s = find(abs(dis) ≤ t)
10:  st = Standard-deviation(s)
11:  if (length(s) > bestSupport) or
    
```

```

(length(s) = bestSupport and st < bestStd) then
12:  bestSupport = length(s)
13:  bestPlane = pl; bestStd = st
14: end if
15: i = i + 1
16: end while
    
```

IV. EXPERIMENTAL RESULTS

In this work as input source for data acquiring we adopt Microsoft Kinect Sensor with OpenNI (Open Natural Interaction) driver which enables communication with RGB and depth cameras of Kinect, and OpenCV (Open Source Computer Vision) library is used for processing depth images.

Data acquisition is done by capturing data from two different scenes, Office Data and Corridor Data as shown in Fig. 3.

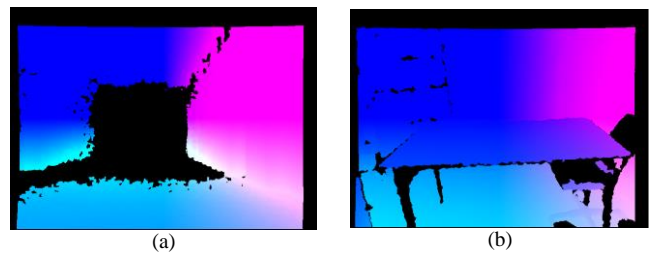


Fig. 3. Original Data Captured by Kinect .(a) Corridor Data, (b) Office Data

A. Camera Calibration

IR camera intrinsic and extrinsic parameters are estimated using Zhang's calibration method [14] and traditional known structure chessboard pattern is used as calibration object as shown in Fig.4, containing 10x7=70 corners, and each square is size of 25.5 mm

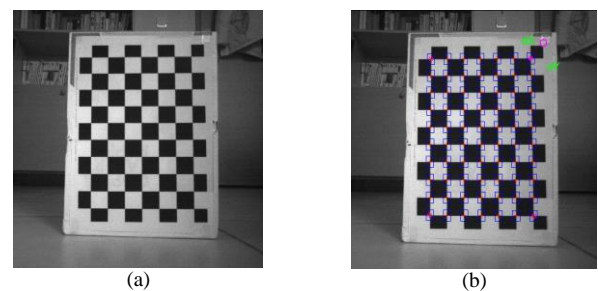


Fig. 4. Camera calibration: (a) chessboard pattern, (b) detected corners.

In Table I and Table II focal length, principal point and distortion coefficients are shown respectively after camera calibration process is done.

TABLE I. FOCAL LENGTH AND PRINCIPAL POINT

fx	fy	cx	cy
----	----	----	----

582.74107	582.74107	319.5	243.5
-----------	-----------	-------	-------

TABLE II. DISTORTION COEFFICIENTS

k1	k2	p1	p2	k3
0	0	0	0	0

B. Hole filling

Two methods are implemented for filling holes. In the first one, as shown in Fig.5, holes are filled by using 10 depth images captured of the same scene without changing the camera view orientation. For 10 depth images average of each pixel is calculated to get the new depth image. The non measured pixels are not taken into calculation.

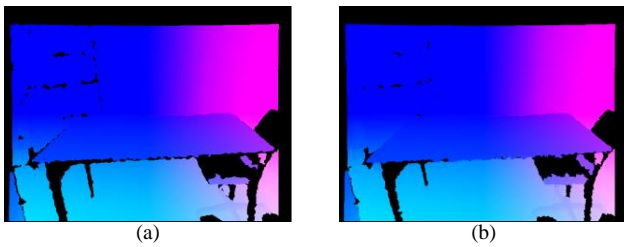


Fig. 5. First method of hole filling, (a) Original Depth images captured by Microsoft Kinect, (b) the resultant filtered depth image using first method

In the second method as shown in Fig.6, 3x3 Median Filter is applied on the resultant image of the first method. By applying such filters on images edge information is lost. However such information is not important to detect plane, applying this filter on depth images does not affects negatively the plane segmentation results.

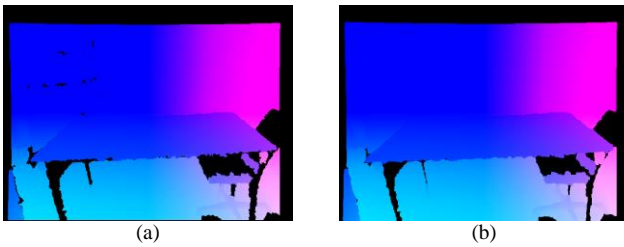


Fig. 6. Second method of hole filling, (a) Left: The resultant filtered depth image with first method, (b) resultant median filtered depth image

C. Depth to point cloud

As it is seen from IR camera calibration results the distortion coefficients are zero. This is because acquiring data using OpenNI driver with Microsoft Kinect sensor returns a processed image. Due to zero distortion it is not necessary to undistort depth images before converting to point cloud data.

The depth data acquired from Kinect using OpenNI driver is in the form of 16-bit 2-D intensity image. Transforming a depth image to point cloud data as shown in Fig.7 using IR camera intrinsic parameters, the focal length and the principal point are required

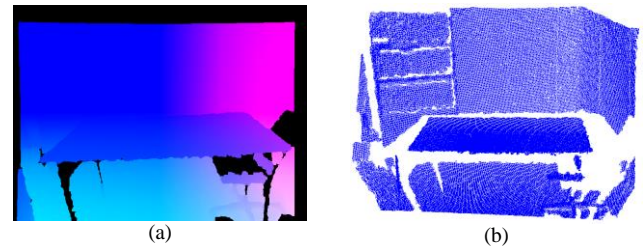


Fig. 7. Converting depth image to point cloud data, (a) Depth Image, (b) Point Cloud Data

D. Removing Non-measured Pixels

The last remaining non-measured pixels are removed from depth image by applying *removeNaNFromPointCloud* filter which is implemented on PCL (Point Cloud Library) [18]

E. Removing Noise

Noise is removed using PCL by applying *StatisticalOutlierRemoval* filter as shown in Fig.8. The best result was achieved by passing these parameters to filter for neighborhood size $k=50$ and $2.5*\sigma$ distance from the mean distance μ .

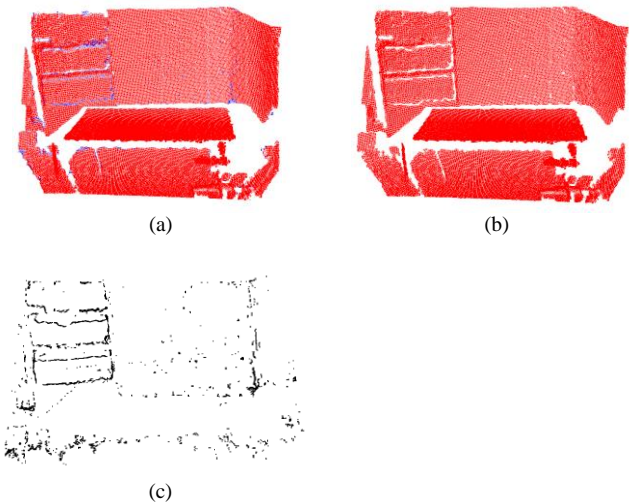


Fig. 8. An example of applying StatisticalOutlierRemoval filter to point cloud data. (a) point cloud data, (b) point cloud data after StatisticalOutlierRemoval operator is applied, (c) noise of point cloud data (outlier).

F. Downsampling

After data is filtered, in order to decrease the process time of plane segmentation voxel grid filter is applied. The size of every voxel is set to 1x1x1cm. Microsoft Kinect sensor produces a depth image with the size of 640x480=307200 points. This is the initial size of point cloud data if no filter is used. Working with such high resolution point clouds for detection planes requires a lot of processing time.

The size of point cloud data after applying all filters is reduced to 73515 points, approximately 24% of the original

data. Reducing the size of point cloud data without losing necessary information will yield very good results also will reduce execution time and cost of CPU power.

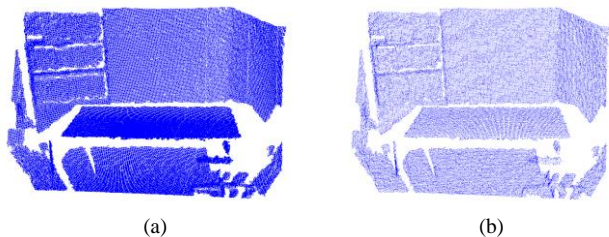


Fig. 9. The result of voxel grid filter. (a) Original data, (b) downsampled data

G. Plane Detection

Plane detection algorithm receives the processed point cloud data as input. Therefore, after all necessary operations are done on point cloud data both data Office Data and Corridor Data are tested by changing the values of input parameter of the algorithm.

RANSAC algorithm implemented on PCL is used for detecting the largest plane. The size of neighborhood k for estimating point normals is chosen 50, Also maximum iteration size equal is chosen as 100.

The algorithm is run several times for each data by changing the values of *NormalDistanceWeight* and *DistanceThreshold* parameters. Evaluation parameter values for the algorithm are as follow: *Normal Distance Weight* = {0.001, 0.01, 0.1, 1} and *DistanceThreshold* = {0.01, 0.05, 0.1, 0.5}.

In total, the algorithm is run 16 times for each data. The best result for Office Data was achieved when *NormalDistanceWeight* = 0.01 and *DistanceThreshold* = 0.1.

Whereas the best result for Corridor Data was achieved when *NormalDistanceWeight* = 0.001 and *DistanceThreshold* = 0.1

As it can be observed from results, for different point cloud data different parameter values have to be used in order to get the best results.

TABLE III. RESULTS OF PLANE SEGMENTATION FOR CORRIDOR DATA

		Distance Threshold			
		0.01	0.05	0.1	0.5
Normal Weighted Distance	0.001				
	0.01				
	0.1				

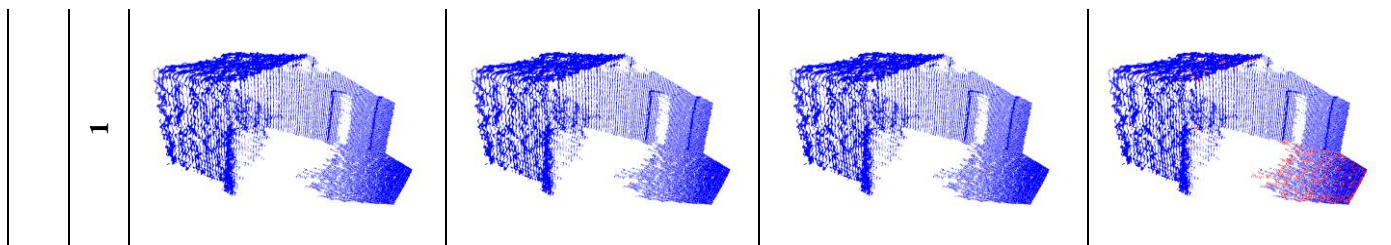


TABLE IV. RESULTS OF PLANE SEGMENTATION FOR OFFICE DATA

		Distance Threshold			
		0.01	0.05	0.1	0.5
Normal Weighted Distance	0.001				
	0.01				
	0.1				
	1				

V. CONCLUSION AND FUTURE WORKS

In this paper an efficient way of data processing and evaluation of *distance threshold* and *normal weighted*

distance parameters for plane detection using RANSAC algorithm is presented. Applying different operations on point cloud data such as hole filling, removing non-measured pixels, removing noise and downsampling the data before

trying to detect the plane yields very good results also reduces execution time and the cost of the CPU power.

From the experimental results it is shown that for different data, different parameter values are needed to get the best results. As a future work, an adaptive method for parameter estimation based on the input point cloud data to achieve robust result is planned.

ACKNOWLEDGEMENTS

The authors would like to thank Research Foundation of the Erciyes University, Kayseri, Turkey for supporting this work under the Grant No. FYL-2015-5601.

REFERENCES

- [1] Y. Guo, M. Bennamoun, F. Sohel, M. Lu, and J. Wan, "3D Object Recognition in Cluttered Scenes with Local Surface Features: A Survey", *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 2013, pp:2270-2287.
- [2] M. Munaro, F. Basso and E. Menegatti, "Tracking people within groups with RGB-D data", 2012 IEEE/RSJ International Conference on, Intelligent Robots and Systems, 2012, Portugal.
- [3] I. Douros and B.F. Buxton, "Three-Dimensional Surface Curvature Estimation using Quadric Surface Patches". *Proc. intern. symp.*, 2002.
- [4] M. Camplani, T. Mantecón, and L. Salgado, "Depth-Color Fusion Strategy for 3-D Scene Modeling With Kinect" *IEEE Transactions On Cybernetics*, 2013, 43, 6, pp:1560-1571.
- [5] F. Endres, J. Hess, N. Engelhard, J. Sturm, D. Cremers, and W. Burgard, "An Evaluation of the RGB-D SLAM System" 2012 IEEE International Conference on Robotics and Automation, RiverCentre, 2012, Saint Paul, Minnesota, USA, pp:1691:1696.
- [6] M.A. Fischler and R.C. Bolles. "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography". *Communications of the ACM*, 1981, 24(6): pp 381–395.
- [7] Z.M. Bi and L. Wangb, "Advances in 3D data acquisition and processing for industrial applications" *Robotics and Computer-Integrated Manufacturing* 26(2010), pp:403–413.
- [8] S. Paulus, J. Behmann, A. K. Mahlein, L. Plümer, and H. Kuhlmann, "Low-Cost 3D Systems: Suitable Tools for Plant Phenotyping", *Sensors* 2014, 14, pp:3001-3018.
- [9] M. Y. Yang, and W. Förstner, "Plane Detection in Point Cloud Data" TR-IGG-P-2010-01, January 25, 2010
- [10] D. Borrmann, J. Elseberg, K. Lingemann, and A. Nüchter, "The 3D Hough Transform for Plane Detection in Point Clouds: A Review and a new Accumulator Design" *3D Research*, pp:1-13.
- [11] H. Paul V C, "Method and Means for Recognizing Complex Patterns," U.S. Patent No. 3069654, 1962.
- [12] H. Woo Yoo, W. Hyun Kim, J. Woo Park, W. Hyong Lee, and M. Jin Chung, "Real-Time Plane Detection Based on Depth Map from Kinect", 2013, 44th International Symposium on Robotics (ISR).
- [13] Y. Suttasupa, A. Sudsang and N. Niparnan "Plane Detection for Kinect Image Sequences", *Proceedings of the 2011 IEEE, International Conference on Robotics and Biomimetics*, December 7-11, 2011, Phuket, Thailand
- [14] Z. Zhang, "A flexible new technique for camera calibration", *IEEE Transactions On Pattern Analysis And Machine Intelligence*, VOL. 22, NO. 11, November 2000
- [15] OpenNI, <https://github.com/OpenNI/OpenNI>, (Online; Accessed: 04 February 2015)
- [16] OpenCV, <http://opencv.org/>, (Online; Accessed: 04 February 2015)
- [17] F. Tarsha-Kurdi, T. Landes, and P. Grussenmeyer, "Extended Ransac Algorithm For Automatic Detection Of Building Roof Planes From

Lidar Data", *The photogrammetric journal of Finland*, 2008, 21 (1), pp.97-109.

[18] PCL, <http://pointclouds.org/>, (Online; Accessed: 04 February 2015)

Instruments of Operator's Active State Identification

Rizun N.

Alfred Nobel University, Dnipropetrovsk
n_fedo@mail.ru

Abstract – In this paper the results of development of the system of tools for identification of operator's active state development were proposed. The types of delay of operator's active state (motor activity) from the point of view of their criticality for the control object were developed. Classification of approaches to correction of each type of delay of operator's active state was designed. The system of qualitative indices of identification of various types of delay was proposed. The functional scheme of technical realization of the device of operator's active state identification using defined approaches and indices was constructed. The methodology of the adaptive correction of the situation of controlling the object by the results of identification of the level of operator's ability to fulfill actions, adequate to the time, was suggested.

Keywords – operator's active state; motor activity; delay; style; classification; device; adaptive correction.

I. INTRODUCTION

In the modern world of human-machine systems, particularly Automated Control Systems, there occurs an abrupt growth of psycho-physiological workload on this system's component – the human operator [1]. In the process of ergative system functioning, under the influence of destabilizing factors in the functional state of operator there occur changes, which can lead to the malfunction or non-fulfillment of regular staff algorithms of operator's activity. One of the most dangerous "failures" of the normative work of such a system is the presence of indications of inadequate operator's response to the regulated actions of the control system (drowsiness, inhibition connected with tiredness, worsening of physical or psychological state), which is identified as person's inability to conduct actions, adequate to the current situation. Such phenomena can occur even in such types of activity, when before each duty (work) period an operator has medical examination of main indices of his physical readiness to start his functional duties.

The number of researches dedicated to this problem is increasing. The biggest difficulty consists in the search of effective criteria for creation of the operator's state control system, which would allow the maximum objective and operative **identification** and **correction** of both the state of operator's working activity and the stable mode of control system operation.

Technical tools of **identification** of operator's state are usually divided into two main groups: with and without contact. Contact method includes the usage of technical control tools, which directly contact the object of examination (a human) – for instance, sensors, fastened in hats, glasses, on earlaps, fingers and wrists etc. The main disadvantage is their constant contact with a person, which may often be an irritant.

In control methods without contact the assessment of state is done distantly, i.e. without tools touching a person. Usually these tools are sensors or computer systems, analyzing eye

movements of an object, change of his body, head or arms position. One of the disadvantages of such methods can be the insufficient informativity, since the information about body or body parts movements does not contain enough data about operator's functional state [1-4].

As an example of existing **methodological** tools of **identification** and correction of the level of operator's working capacity we can consider the following approaches:

- terminal supply of sound signals, the deactivation of which is controlled by a person, or the supply of alarm signal in case of absence of operator's reaction to this signal [5];
- definition of the levels of wakefulness and sleep stages, including the stage of drowsiness in the process of professional activity with the help of shifting the point of pressure of a body to the force-torque chair and returning the operator from sleep to the wakeful state or from drowsiness with the help of controlling signals [6];
- fixation of body's shifting by means of analyzing the level of deformation of the plate under operator's seat and alarm about operator leaving the "working" state [7];
- control of the level of emotional tension by means of revealing the abilities of an operator to perceive selective typical doubled information [8];
- registration of sequence of operator's actions, change of speed of his reaction to irritants and comparison of actual actions with the optimal model of operator's behavior in a particular situation [9];
- test of operator's professional suitability by means of comparing the time of operator's actual reactions to the external factors and the operation of vehicle with the standard one. The comparison is conducted by n levels of complexity; in case of malfunctioning or operator's belated reaction in the number, equal to the number of complexity level plus one, the testing is continued on the next, upper level of complexity [10];
- application of electrical sensor, placed in the zone where operator's fingers influence its parameters, and

activation of the alarm signal in case when within the set time period the information about operator motor activity is absent [11];

- designing on the basis of Operator Functional State pattern recognition methods an adaptive aiding system either to remind the operator or to reduce the task load during the period of excessive mental workload, with an aim to enhance the overall system performance [12, 13, 14, 15];

- adaptive fuzzy model linking heart-rate variability and task load index with the subjects' optimal performance via a series of experiments involving process control tasks simulated on an automation-enhanced Cabin Air Management System [18, 19, 20].

Even considering the limited list of given examples of methodology of identification and correction of level of operator's active state, we can point out the following disadvantages:

- improvement of only the algorithms of identification of non-typical (non-regulated) operator's behavior with the further activation of signal about anomaly detection [5, 6, 7, 9, 11, 18, 19] without realization of complex approach to the possibility of correction of the revealed behavior deflections. This approach can be optimal only in case of heightened danger of failure situation occurrence, which does not give additional time to conduct a more thorough situation analysis;

- expansion of the area of application of the methodology of controlling operator's active state by means of using technical tools and algorithms of adaptive analysis of the degree of deflection from human's regulated standard motor activity [8, 10, 12, 13] with the use of test control, doubled information perception etc.

However, each of the mentioned **methodologies** possesses **limitations** by:

- the type of the defined deflection (emotional, visual perception, information distortion);

- the classification of types of operator's anomaly behavior (the degree of closeness to sleep, the degree of emotional overload etc.);

- the time of application of the adaptive correction methodology (for instance, not in the process of work but before the beginning of operator's shift);

- the type of influencing controlling signals (doubled information, sound irritants, semantic tests);

- the type of operator's work place (vehicle, autonomous stations, boiler houses etc.).

Thus, we can make a conclusion that there exists no single methodology of **identification** and **correction** of operator's working activity.

The authors set the objective of *providing the required level of quality of control and regulation of the parameters of Automated Systems* due to the development the system of the instruments, which contains:

1. Classification of types of delay of operator's active state (motor activity) from the point of view of their **criticality for the control object**.

2. Classification of approaches of **correction** of each type of delay of operator's active state.

3. Qualitative **indices** of identification of various types of delay of operator's active state.

4. Functional **scheme** of technical realization of the **device** of operator's active state identification using defined approaches and indices.

5. Complex **methodology** of:

- identification and obtainment of quantitative indices of specific characteristics of an operator concerning his ability to *maintain the active working state during the shift*,

- adaptive *correction of the situation of controlling* the object by the results of identification of the level of operator's ability to fulfill actions, adequate to the time.

II. BASIC CONCEPTS OF IDENTIFICATION OF OPERATOR'S ACTIVE STATE

The suggested system of the instruments for identification of operator's active state is based on the following concepts, developed by the authors:

Concept 1. Delays in the active state of an operator (his/her motor activity) of the Automated Control System can be divided into 3 main types:

- **working (type A)** – are mostly connected with the discrete character of the control process, done by the operator;

- **non-critical (type B)** – can be connected with an accidental shift of operator's attention to the extraneous (which are not restricted by the job description) objects and actions;

- **critical (type C)** – can be connected with the operator's loss of attention and ability to carry out conscious professional actions, which can be dangerous for the process of object control.

Concept 2. **Correction** of each of the three defined types of delay of the motor activity of an operator, which may appear, can be fulfilled in accordance with the following algorithms:

- **regulated** – by means of an *independent* (without applying the identification tool) restoration of operator's motor activity in connection with the appearance of the next *regulated* operation of controlling an object;

- **test correction** – by means of the *automated* (with the help of the identification tool) restoration of operator's motor activity by using the *test* algorithm of *shifting operator's attention* from current to special non-typical tasks of decision-making and decision-realization;

- **failure correction** – by means of the *automatic* (with the help of the identification tool) interference into the process of control by indication of the *signal* of appearance of a *critical situation* or by *switching* the control system into the *automatic* regulation mode.

Concept 3. **Identification** of the defined types of the delay of motor activity of an operator of the Automated Control System is suggested to be carried out with applying the following **indices**:

– **regulated** average time of the possible absence of operator’s motor activity T_k (defined on the basis of statistical data – results of observations of the etalon operator’s work during one shift);

– **non-regulated** test time T_{test} that consists of the **regulated** average time T_k and time for the **test** correction of operator’s motor activity T_r and which totally must not exceed the *maximum non-critical* regulated time T_{max} of the possible absence of operator’s motor activity:

$$T_{test} = T_k + T_r, \quad (1)$$

$$T_{test} \leq T_{max}. \quad (2)$$

– **failure** time T_{fail} , which consists of the **regulated** average time T_k for the test correction of operator’s motor activity T_r and the time for decision-making about the choice of algorithm of failure correction T_{alarm} , and which totally must not exceed the *maximum critical* regulated time T_{A_max} , within which the Automated System can work without the operation of process control.

$$T_{fail} = T_k + T_r + T_{alarm}, \quad (3)$$

$$T_{fail} \leq T_{A_max}. \quad (4)$$

III. MAIN SCIENTIFIC SOLUTIONS FOR THE OPERATOR’S ACTIVE STATE IDENTIFICATION

A. Functional Scheme of the Device

To solve the technical part of the complex task of system of the instruments for identification of operator’s active state development we suggest the device, which includes (figure 1):

1 – block of the electrical Sensor of operator’s motor activity (for example, infrared scanning laser); placed in the zone, where the position of operator’s hands can influence its parameters;

2 – block of the Timer (for instance, a specified chip, which is activated by electrical signals);

3 – block of the Programmer;

4 – Alarm block;

5 – block of Signal Elements, which is placed in the range of operator’s vision and is supplemented by five signal elements, which set the position of the manual manipulator: left, right, up, zero, down.

6 – block of Control Keys;

7 – block of the Manual Manipulator with five finite positions: left, right, up, zero, down, which are placed in the zone of influence of operator’s free hand fingers;

8 – Comparator block;

9 – Accumulator block.

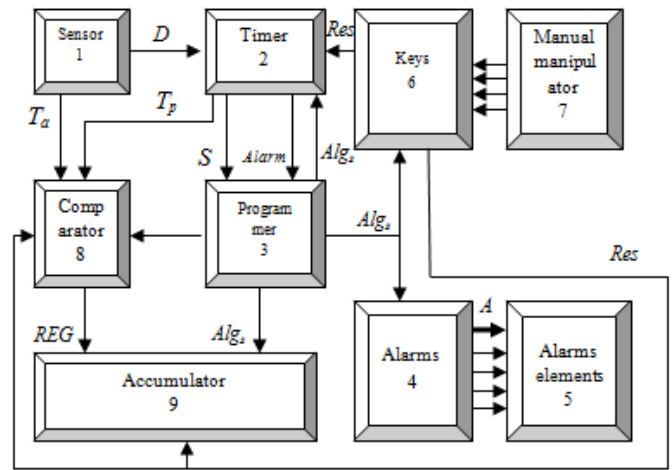


Fig. 1. Functional Scheme of the Device for Identification of the Operator’s Active State

B. Algorithm of the Complex Methodology Identification

On the basis of the proposed concepts and technical solutions of the research goal, the **algorithm** of the suggested methodology of operator’s active state identification is propose. It includes the following steps:

1. To ensure work of the device before the process of identification the preparatory works are carried out:

- conduction of experiments and calculation of indices of the average **regulated** time T_k of the possible absence of operator’s motor activity; **non-critical regulated** time T_{max} of the possible absence of operator’s motor activity; **maximum critical regulated** time T_{A_max} , within which the automated system can work without the operation of process control; optimal time T_{opt} for the recognition and perception of information from the signal element, and also directly the conduction of commutation by the manual manipulator.

– keeping these indices in the blocks of the timer, programmer and comparator.

2. In the Timer the initial recording time is set – **regulated average** time T_k of the possible absence of operator’s motor activity

3. Conduction of each (i) regulated professional action (movement) of an operator is accompanied with the transfer of the time signal D from the Sensor of operator’s movement to the timer.

4. Within the time from the moment of transfer of the previous signal about professional actions (movements) of an operator the timer conducts the countdown (starting from zero) with the constant **control** of the absence of exceeding of the regulated average time T_k by the current time in the Timer

$T_p(i)$:

$$T_p(i) \leq T_k. \quad (5)$$

5. In case when the moment of arrival of the signal $T_p(i)$ from the Sensor of operator's movement to the timer conforms the condition (5), the Comparator identifies the (*i*) current state of the operator with the type of **working** delay of operator's activity (type *A*).

6. At the same time, the *current* value of delay time, saved in the Timer, is nullified $T_p(i) = 0$ and the timer starts the new countdown.

Restoration of operator's motor activity is carried out following the **regulated** algorithm – i.e. by means of an *independent* restoration of motor activity connected with the appearance of the next *regulated* operation of controlling an object.

7. Information about each (*i*) current state of an operator goes from the movements Sensor and the Timer, is saved in the Accumulator of the device in the format of tuple *REG*, which keeps *quantitative* characteristics of the delay of operator's motor activity – the period of time before the action $T_p(i)$ and the action time $T_a(i)$, as well the *qualitative characteristic* – the type of delay of motor operator's activity:

$$REG = \langle T_a(i), T_p(i), A \rangle. \quad (6)$$

8. In case the time signal *D* from the movement Sensor does not come by the moment when the *current* time in the timer $T_p(i)$ equals to the regulated average time T_k , the timer sends signal *D* to the Programmer, informing about the necessity to use the algorithm of **test correction** of the delay of operator's motor (working) activity.

9. For this purpose the Programmer, in accordance with the set in it programs, forms (chooses) the algorithm Alg_s of the **test correction** within the time T_r , which presupposes transmission of the signal to the Alarm block concerning the sequence and time of activation of the signal elements *Left, Right, Zero, Down, Up* within the time, which does not exceed T_r (with the condition $T_k + T_r \leq T_{max}$):

$$ALG_s = \{Seq, \langle t_1, t_2, t_3, t_4, t_5 \rangle\}. \quad (7)$$

The difference between the time moments of signal elements activation must not exceed T_{opt} :

$$t_i - t_{i-1} \leq T_{opt}. \quad (8)$$

10. The Programmer also sends the signal Alg_s to the Timer when the algorithm starts and it begins the countdown with the constant **control** of absence of the exceeding of the *maximum non-critical* regulated time T_{max} by the current time in the Timer $T_p(i)$:

$$T_p(i) \leq T_{max}. \quad (9)$$

11. The same signal Alg_s is sent by the Programmer also to the block of Control keys.

12. In case the operator conducts the **correct** commutation – i.e. the concurrency of the complex of indices of sequence *Seq* and the commutation time:

$$Res = \{Seq, \langle t_1 + t_{opt}, t_2 + t_{opt}, t_3 + t_{opt}, t_4 + t_{opt}, t_5 + t_{opt} \rangle\} \quad (10)$$

– the signal *Res* from the Manual Manipulator through the open proper key of the block gets to the Timer, after that the current value in the Timer is nullified $T_p(i) = 0$ and the timer starts the new countdown;

– the signal *Res* from the Manual Manipulator and signal Alg_s from the programmer get to the Comparator, which, if the condition (9) is confirmed, identifies the (*i*) state of the operator by the type of the **non-critical** delay of operator's activity (type *B*);

– the signal *Res* from the Manual Manipulator gets to the Accumulator and is saved there in the format of tuple *REG*, which keeps *quantitative* characteristics: of duration of the *non-critical* delay of operator's motor activity $T_p(i) = T_a(i) + T_k$; the algorithm of test correction and the moment of time, when it will be successfully conducted $T_r(i) = T_p(i) + T_r$; *qualitative* characteristics – the type of delay of operator's motor activity:

$$REG = \langle T_p(i), T_r(i), B \rangle. \quad (11)$$

Thus the fulfillment of the **correct** commutation by the operator proves the fact that the operator possesses the ability to conduct conscious adequate professional actions, which was identified and corrected by means of the automated *test* algorithm of *shifting operator's attention* from current to non-current tasks of decision-making and decision-realization.

13. In case the operator conducts the **incorrect** commutation – i.e. the non-concurrency of the complex of indices of the set sequence *Seq* and/or the commutation time (7):

– the signal from the Manual Manipulator does not reach the Timer, that is why the timer continues the countdown with the constant **control** of the fact that the current time in the timer $T_p(i)$ does not exceed the *maximum critical regulated* time T_{A-max} , within which the Automated System can work without conducting the process of operation regulation:

$$T_p(i) \leq T_{A-max}; \quad (12)$$

– the timer send the signal *Alarm* to the Programmer, meaning that it is necessary to use the **alarm correction** – by means of the automatic interference into the object control;

– in the Programmer (according to the installed programs) the algorithm Alg_s is formed (chosen); it presupposes the transmission of the signal to the Alarm block with the help of the signal element, informing about the

critical situation within the time period that does not exceed T_{alarm} ;

– Alg_s signal gets from the programmer to the **comparator**, which in case of confirmation of the condition (6) identifies the (*i*) current state of an operator by the type of a **critical** delay of operator's activity (type C);

– on the basis of signals from the Timer and the programmer Alg_s , the Accumulator stores information REG , which keeps *quantitative* characteristics of the duration of *failure* delay of operator's motor activity $T_p(i) = T_a(i-1) + T_{fail}$ and *qualitative* characteristics – the type of delay of operator's motor activity:

$$REG = \langle T_p(i), C \rangle. \quad (13)$$

14. Parameters of the algorithm of **test** correction can vary depending on:

– the number of idle moments in operator's motor activity during the current shift K_v ;

– the average number of idle moments in operator's motor activity during the shifts \bar{K}_v ;

– the current time of a shift which corresponds to the five types of operator's functional state, notably:

– initial reaction (I) – a short-term decrease of the actual level of confidence and accuracy of operator's actions;

– hyper-compensation (II) and compensation (III) – gradual increase and stabilization of indices of confidence and accuracy of operator's professional activity to his individual actual level (the period of norm maintenance);

– sub-compensation (IV) and de-compensation (V) – decrease of operator's normal level of confidence and accuracy of professional activity, connected mainly with the tiredness.

15. In parameters of the algorithm of **failure** correction the variants of realization of operation of *switching* the system of object control into the *automatic* mode can also be presupposed.

Technical results of the suggested system of the instruments for identification of operator's active state consists in *the increase of quality of regulating the work parameters of automated systems of control* by means of expanding the area of application of the tools of operator's active state identification within the shift with the aim of giving the possibilities of *control* and *correction* of situations, which happen in the process of control, by the results of analyzing the type of delay of operator's motor activity and his ability to conduct conscious actions.

The tables 1 and 2 include results of the imitation experiment with the identification of operator's active state according to the given functional scheme (figure 1).

TABLE I. SOURCE DATA OF TIMETABLE (C)

Regulated average time of the possible absence of operator's motor activity T_k	0:00:10
Time for the test correction of operator's motor activity T_r	0:00:15
Maximum non-critical regulated time of the possible absence of operator's motor activity T_{max}	0:00:30
Time for decision-making concerning the choice of algorithm of failure situation correction T_{alarm}	0:00:10
Maximum critical regulated time within which the automated system can work without the operations of process control T_{A-max}	0:00:50
Optimal time for recognition and perception of information from the signal element and also the direct conduction of commutations by the manual manipulator T_{opt}	0:00:03

TABLE II. RESULTS OF THE IMITATION EXPERIMENT OF OPERATOR'S ACTIVE STATE IDENTIFICATION

<i>i</i>	$T_a(i)$	T_r	Algorithm Alg_s of test correction					Time period of successful test fulfillment $T_i(i)$	$T_p(i)$	Total time	Value of the timer signal	Type of motor activity delay	Signal about the critical situation occurrence
			$t_{1\ LEFT}$	$t_{2\ RIGHT}$	$t_{3\ ZERO}$	$t_{4\ TOP}$	$t_{5\ DOWN}$						
1	0:00:03	-	-	-	-	-	-	-	0:00:03	0:00:03	D	A	-
2	0:00:06	-	-	-	-	-	-	-	0:00:06	0:00:09	D	A	-
3	-	0:00:09	0:00:00	-	0:00:03	0:00:06	-	0:00:09	0:00:19	0:00:37	S	B	-
4	0:00:05	-	-	-	-	-	-	-	0:00:05	0:00:42	D	A	-
5	0:00:06	-	-	-	-	-	-	-	0:00:06	0:00:48	D	A	-
6	0:00:06	-	-	-	-	-	-	-	0:00:06	0:00:54	D	A	-
7	0:00:04	-	-	-	-	-	-	-	0:00:04	0:00:58	D	A	-
8	-	0:00:15	0:00:00	0:00:03	-	0:00:06	0:00:09	0:00:12	0:00:22	0:01:35	S	B	-
9	0:00:06	-	-	-	-	-	-	-	0:00:06	0:01:41	D	A	-
10	-	0:00:15	0:00:00	0:00:06	0:00:03	0:00:12	0:00:09	-	0:00:15	0:01:56	Alarm	C	A

IV. CONCLUSIONS

Therefore, the suggested system of the instruments for identification of operator's active state and the methodology of its application allows to:

1. Provide the increase of **quality** of control and regulation of parameters of Automated Systems by means of introducing algorithm and tools:

– of identification of the type of motor activity delay of an operator of automated control system;

– of adaptive correction of non-critical and critical types of delay of operator's motor (working) activity by means of test and failure correction.

2. Increase the **efficiency** of operator's work by means of implementing the technology of maintaining the active state of work within the shift as well as by means of prompt interference into the process of control in case of defining the critical situation of the absence of operator's possibility to conduct actions, adequate to the time requirements.

REFERENCES

- [1] N. Mikhailov, I. Fudimov, "The results of the development of instruments for objective control of the waking state" in *Electronic scientific journal "Engineering Gazette Don"*, № 2, 2009, pp.201-205.
- [2] N. Ponomarev, "Automated control system of the functional state of the driver of the vehicle" in *Transport of the Russian Federation*, № 2 (27) 2010, pp. 22-23.
- [3] A. Whitlock, "Driver Vigilance Devices: Systems Review" in *Railway Safety. Quintec*, 2002, pp.99-103.
- [4] A. Williamson, T. Chamberlain, "Review of on-road driver fatigue monitoring devices" in *University of New South Wales*, April, 2005, pp.43-52.
- [5] B. Semenov, "A method for controlling health staff on duty and alert control center and a device for carrying out the method", the patent RU 2405208.
- [6] E. Stadnikov, A. Sliva, N. Stadnikova., V. Kostecki., "The method for controlling and managing the functional state of the operator and device implementation", the patent for an invention, RU 2417053
- [7] C. Borodin, "Device for control and management of the functional state of the operator". The patent for an invention, RU 1708303.
- [8] V. Savchenko, "Device control the functional state of the human operator", the patent for an invention, RU 2020871.
- [9] B. Malich, "The control system functional state of the human operator in the performance of his duties as a stationary workstation", the patent for an invention, RU 2199272.
- [10] V. Suholitko, "A method for controlling the professional suitability of the operator", the patent for an invention, RU 2199272.
- [11] B. Gerasika, "Pristriy for the active control will become the operator", the patent for an invention, UA 58957.
- [12] Hockey GRJ, "Operator functional state: the assessment and prediction of human performance degradation in complex tasks" in *US: IOS Press*, 2003, pp. 193–200.
- [13] Hockey GRJ, Wastell DG, Sauer J., "Effects of sleep deprivation and user-interface on complex performance: a multilevel analysis of compensatory control" in *Hum Factors*, 1998; 40, doi: 10.1518/001872098779480479, pp. 233–253.
- [14] Jian-Hua Zhang, Xiao-Di Peng, Hua Liu, Jörg Raisch, and Ru-Bin Wang, "Classifying human operator functional state based on electrophysiological and performance measures and fuzzy clustering method" in *Cogn Neurodyn*, v.7(6); 2013 Dec, PMC3825145, pp. 190-199.
- [15] Nickel P, Hockey G.R, Roberts A.C., Roberts M.H., "Markers of high risk operator functional state in adaptive control of process automation" in *Proceedings of IEA*, 2006, pp. 304-312.
- [16] Maurice de Montmollin, "Analysis and Models of Operators Activities in Complex Natural Life Environments" in *PRODUCTION, N° Especial*, 2000, pp. 29-42.
- [17] Tattersall A.J, Hockey G.R., "Level of operator control and changes in heart rate variability during simulated flight maintenance" in *Hum Factors*, 1995, 37, doi: 10.1518/001872095778995517, pp. 682–698.
- [18] Ching-Hua Ting, Mahfouf M.; Nassef A., "Real-Time Adaptive Automation System Based on Identification of Operator Functional State in Simulated Process Control Operations" in *Systems, Man and Cybernetics*, Volume:40 Issue:2, 2008, pp. 203-210.
- [19] Gevins, A., & Smith, M.E., "Neurophysiological measures of cognitiveworkload during human-computer interaction" in *Theoretical Issues in Ergonomics Science*, 4, 2003, pp.113-131.
- [20] Boel M., & Daniellou F. "Elements of process control operator's reasoning: Activity planning and system and process response-time. In *Ergonomics Problems in Process Operations*" in *European Federation of Chemical Engineering Publications*, (Series N38): The Institute of Chemical Engineering and Pergamon Press, 1984, pp.332-340.

Multimedia and Its Applications

How Infographic should be evaluated?

Waralak V. Siricharoen
School of Science and Technology
University of the Thai Chamber of Commerce
Bangkok, Thailand
Waralak_von@utcc.ac.th

Nattanun Siricharoen
Faculty of Communication Arts
Huachiew Chalermprakiet University
Samutprakarn, Thailand
Nattanun2004@yahoo.com

Abstract–With a little time and too much information to learn, Infographics were used to support as data visualization. Visual communication is communication through visual aids. Visual communication is more effective than reading and hearing information. There are so many infographics for everything; however, how can we know which one has a good or not so good design. The aim of the paper is to clarify the evaluation approach of infographic in many aspects and methods. The paper presents the definition and the uses of infographic in the introduction section. The types of infographics are mentioned with the examples. The trendy multimedia and interactive infographic are introduced. The most important part of the paper is the evaluation approach, which will be addressed with the questions and discussion for infographic developer. The recommendation for infographic designing will be concluded in the last section of the paper.

Keywords- *Infographic, Data Visualization, Information Visualization, Infographic Evaluation*

I. INTRODUCTION

Infographic is data visualizations that present complex information quickly and clearly [1]. As mentioned by Visual.ly, data visualization includes signs, photos, maps, graphics and charts, it presents complex data. The infographic is part of data visualization. The foundation of infographics is composed of three major parts. They are Visual, Content and Knowledge. Visual representations of data, information, and/or knowledge are:

- (1) Visual elements – colors, graphics, signs, icons, maps, etc.
- (2) Content elements –facts, statistics, texts, references, time frames, etc.
- (3) Knowledge –conclusion to express the stories or messages [1].

The infographic design is very significant. The designing process of infographic can help understand and implement the principles of the designs better than designing the web or documents as well. Nevertheless, graphics could fast persuade readers to disregard the article [2]. Now almost everyone is using infographics, such as companies, educators, non-profits organization, etc. The reasons why we need infographics are to communicate a message, to present large amount of information in a compressed and easy way to understand, to expose the data, to determine cause-effect relations, and to classify relationships among data, and to observe changes or trends in data [3]. The tools help integrate these three main components to form nice and well-designed infographics [2]. Importance of infographics are making information more appealing, showing valuable ideas, attention-grabbing, easier to understand, being more persuasive, memorable, easily relay information [4]. The tools help integrate these

three main components (Visual, Contents and Knowledge) to form nice and well-designed infographics. The word “infographic” according to the book of Krum [23] mentioned (based on the data from Google Insights for Search) shows that the last two years (2010-2012) the clearly growth in searching for “infographic” term has been recognized.

II. TYPES OF INFOGRAPHICS

In general, there is the accepted standard in the medium towards exploring new formats based on web mash-ups and data visualization, but not often they aim to make up a space for public debate that provides readers more than just only one platform. Arabic numerals are preferable in infographics; the heading of table should be underlined and centered above them. Human mind can recognize visual information much more successful. With today’s technology, infographics can also be transformed to animated images for the website version [5]. There are many articles and papers mentioned the types of infographics. They are the followings: Ashton [6] said that at least we had many types of infographics coming in all shapes and sizes [7]. The followings are the main types of infographics:

- The Visual Article (Figure 1)
- The Flow chart (Figure 2)
- Useful attraction (Figure 3)
- Number, The Timeline (Figure 4)
- Data Visualization(statistical based) (Figure 5)
- The Compare & Contrast (“VS: Versus”) (Figure 6)
- The Photo (Figure 7)
- How-to (process oriented) (Figure 8)
- Research Results(Figure 9)
- “Did-You-Know?” (Figure 10)



Figure 1. Example of visual story infographic¹(Titanic)



Figure 2. Example of the flow chart infographic²(How Affiliate Marketing Works)



Figure 3. Example of useful attraction infographic³(Travel Like an Athlete)



Figure 4. Example of number, the Timeline infographic⁴(Google Search Timeline)

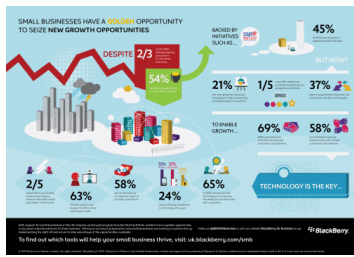


Figure 5. Example of data visualization infographic⁵(Technology is the key)

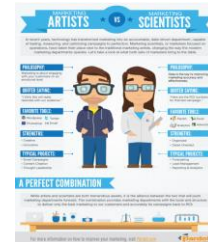


Figure 6. Example of VS infographic⁶



Figure 7. Example of the photo infographic⁷(Wedding Infographic)



Figure 8. Example of how-to (process oriented) infographic⁸(Tie Instruction)



Figure 9. Example of Research Results infographic⁹(Top Employers South Africa 2014)



Figure 10. Example of "Did-You-Know?" infographic¹⁰(Winter Tyre Statistics)

¹ http://thumbnails-visually.netdna-ssl.com/titanic_521df2c4be587_w1500.png

² <http://designwebkit.com/wp-content/uploads/2011/10/Affiliate-Marketing-flowchart.jpg>

³ http://www.jaunted.com/files/6193/Travel_Like_an_Athlete_COMPLETE.jpg

⁴ http://1.bp.blogspot.com/-JMta3XWwP1M/UkSAUDID_ZI/AAAAAAAAAqs/H_JBg0HOaa8/s1600/Screen+Shot+2013-09-26+at+11.40.10+AM.png

⁵ <http://bluehatmarketing.com/wp-content/uploads/2013/02/infographics-and-data-visualisation-image1.png>

⁶ <http://www.pardot.com/wp-content/uploads/2013/01/Marketing-Scientists-vs-Marketing-Artists.png>

⁷ http://fc03.deviantart.net/fs70/i/2013/142/f/b/wedding_infographic_design_by_darkstalkerr-d66603e.jpg

⁸ <http://mlhart.files.wordpress.com/2011/02/tie.jpg>

⁹ <http://www.top-employers.com/PageFiles/6501/Top%20Employers%20South%20Africa%20Research%20Results%202013%20Infographic.png>

¹⁰ http://thumbnails-visually.netdna-ssl.com/did-you-know_50ffd4e64bf9d_w1500.jpg

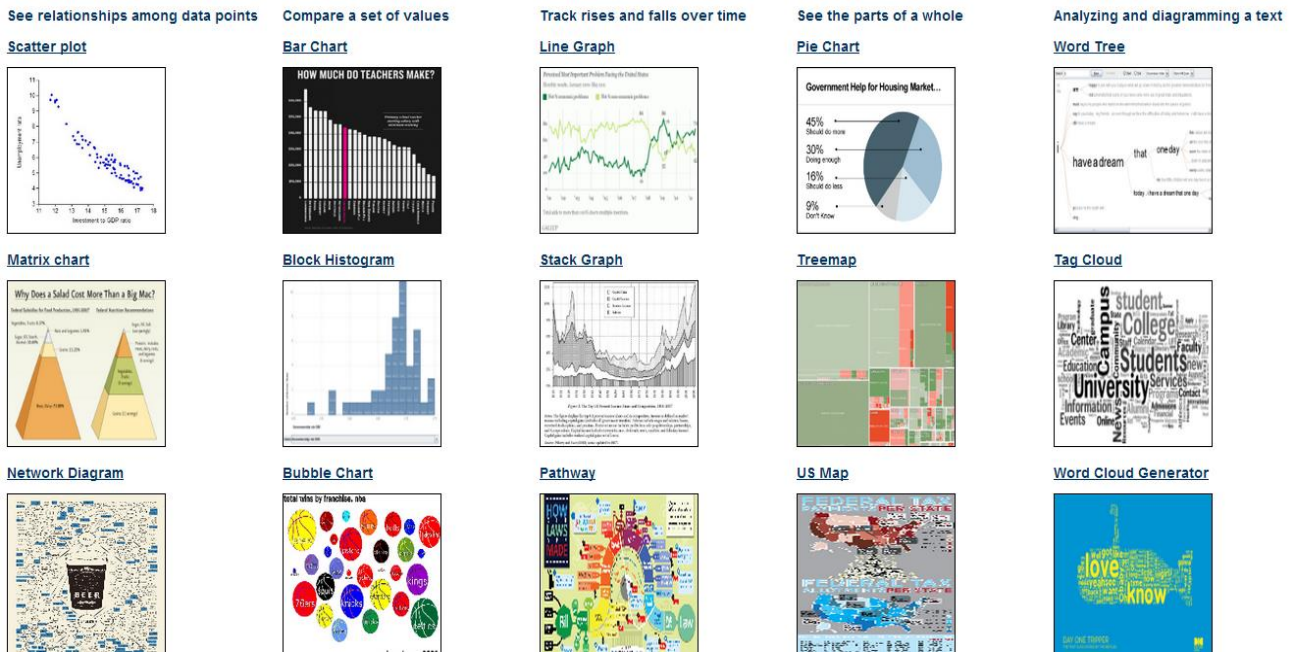


Figure 11. The way to present data in infographic [22]

To create an infographic, one needs the creativity and aesthetic. The infographic creators sometimes have to act as artists. Because the main purpose of infographic is to make people summarize information quickly in the world full of flooding information. Infographic is not made for entertaining only but it should conclude the main ideas/products/services. The infographic creators should consider the foremost structure, accuracy, reliability, depth, and functionality and then think about decoration. The infographic creators must think about whether the decoration is required at all or what type of infographic would appropriate for presenting the main idea. This is not that infographics should not be beautiful. On the other hand, if infographic has been made in order to draw readers and make them feel attracted in contents. If decoration can be the obstacle for readers to understand the main idea of the infographic, possibly the aesthetic or elegance may not be needed.

Linda [8] said in her presentation on website slideshare.com that the infographics need analyzing, evaluation, and creation. To present the huge information in a large set of data, graphs and statistical chart or geographical context to a story with a map need to be applied. These are data visualization tools. The pattern is hidden behind the complexity of tabular forms of numbers and data. The pattern became visible when a human brain can recognize it. The human brain can easily remember things in picture better than texts. The well-built and reasonable presentation of information, graphics, charts, maps and colors are what make infographics noticeable and attract the audience successfully [9]. In Figure 11, it shows the guideline of how creators create the type of graphs or

charts to present or emphasize the different situations. For example, if the creator would like to see the relationships among the data points, the suitable graph or chart type would scatter plot, matrix chart, and network diagram. Line graph, stack graph, and pathway should be used for tracking rises and falls over time. Bar chart, bubble chart, and block histogram are tools for comparing different values with correctness, and a choropleth map is a tool for distinguishing the comparative occurrence of phenomena in an area and can illustrate the piece in the whole map. Good graphics are not just displays; audience can extract information from, but devices to explore information with [10].

III INTERACTIVE, MULTIMEDIA AND VIDEO INFOGRAPHICS

Creating the graphical representation of complex datasets makes it easier to be understood; this is called data visualizations and infographics. One original infographic project is Paris Metro by Harry Beck [14]. The meaning of Harry Beck's map design in 1933 (Figure 12, 13) is widely familiar as one of the Graphic and Information Designs, and became a main persuasion to the underground maps of the world. In year 2009, Mark Ovenden wrote the book called, "Transit Maps of the World", in which he mentioned the map of Harry Beck.



Figure 12.& 13. The old maps used as the infographic.

Nowadays everything is interactive; including interactive infographics which allow the users to interact with the information. They need designers and programmers to create and allow the users to explore the information themselves. One of well-established interactive infographics has been created by news organization like CNN. For example, the CNN Ecosphere uses WebGL to create data visualization; in addition, since the social networking is very well known for everyone, the interactive infographic can help spread the infographic.

The effort for creating good interactive infographics is greatly increasing, nowadays many web-based data visualization tools are making it easier to create interactive infographics, but they still require great effort especially when multiple datasets are being compared [15]. The information can change over time and that can create the interactivities in infographics. The interesting Interactive/Video Infographics sites are as follows:

- <http://www.dipity.com/> (create an interactive online timeline)
- <http://www.tableausoftware.com/public/community> (live metric tracking)
- http://news.bbc.co.uk/2/hi/in_depth/interactives/default.stm
- <http://www.coolinfographics.com/blog/tag/video>
- <http://youtu.be/d0zION8xjBM>
- <http://vimeo.com/9602282>

More and more interactive infographics have been used widely, they are very catchy and improving user-experiences [11]. Some of the most engaging infographics are actually interactive pieces created in Flash or HTML5. More detailed information can be discovered by clicking any words or pictures. The interactivity of the feature is joined with the immediacy and interest that the topic generated combined to create a very successful piece of link attraction [12].

The multimedia infographic allows some interactivity, and the users have an ability to browse the infographic which offers a virtual connection to the content. New technologies can and do encourage designers and visual communicators. Video infographic offers another engaging format for infographics [13]. This will include imaging multiple panes of static or animated infographics with a voice over to tell a story. The impact is more. The moving images are very attractive and they prevent our

concentration to deviate [12]. The advantages of animated infographics are: The benefits of animated infographics¹¹: They are static but particularly ground-breaking and eyes catchy. They facilitate to clarify the difficult and complex idea for examples in medical and health issues. They can be plugged to the blogs and website where static infographics are submitted. In video infographic, the animated text can move towards flying in and out. It will easily receive the main message together with adding to that visuals, background music and sounds. Moreover, the font and its size are very significant. The text must be very clear to the viewers' eyes and the font has fair readability. Animated interactive infographics will happen to common by using Flash or HTML5 in websites; it makes infographics much more interesting same as video infographics.

III. CREATING INFOGRAPHICS

The first way to create an infographic is to hire a designer or agency to create it. This is to make sure that the infographic is visually independent of others and representatives of the brand. Nevertheless, it can be expensive. In the old days we could create infographic using paper, pen, pencils, makers, and rulers. Alternatively, today we use computer software to build nice infographic, which is both faster and easier and even more attractive and multicolored [16]. A different way is to find the suitable websites that offer portfolios or templates of infographic, or information on hiring designers. There are two methods in order to create the proper infographic [1].

1. Build entirely online with infographic website (such as www.visual.ly and www.ease.ly)

- Advantages: easy, quicker, graphics and creation tools provided for you, publish and share
- Disadvantages: limited data input, limited template and design choices, may not be higher for printing, maybe restricted to their website

2. Use image editing software to build infographics (MS Excel, MS PowerPoint or Publisher, Photoshop, and other open source software, such as Paint.Net, Gimp, Inscape, Photoshop Express, Pixlr, Sumopaint, Creately) and also there are many Apple and Android application (Apps) to create the infographic. Then host it online

- Advantages: more design freedom, build it high-resolution for print, use it/output it many formats, host it online easily
- Disadvantages: more work, requires a little knowledge of image editing/design principles, find sources for hosting/sharing

Nonetheless, if we need to build the infographic for specific purpose, here is the first start. Krauss suggested a very useful nine-step process to create the infographic [1] as follows:

¹¹ www.inforgraphicdesignteam.com

1. Gather your data (need numbers from more than once recourses)
2. Determine your purpose
3. Plan your infographic (outline or flow chart)
4. Start laying out your plan with software or an online tool (graphics or photos)
5. Evaluate your data
6. Find the best way in a visual representation (type of charts, and cite data properly)
7. Apply a color scheme & choose fonts
8. Step back and evaluate it, get feedback and edit
9. Caution about the copyright, cite your sources for data, and don't use any image off the web.

There are a number of tools available for creating infographics. Some are more user-friendly than others. Several sites that can be used to visualize information. An incomplete sentence Because of the ability to capture attention and convey information in a straightforward manner, infogr.ams have exploded in use and popularity. Students must be able to understand and analyze information presented in this manner. Production of infographics also demonstrates a student's expertise of advanced cognitive skills, technical skills, and familiarity with varied literacies. The widespread use of infographics arises and with emerging technologies influencing the definition and requirements of "literacy" for current and future students and teachers. They need to embrace new forms of presenting information and use it efficiently.

Things that have to be taken to consideration for the infographic are as follows:

1. What are infographic? definition, purpose, types
2. Are there any standards for infographics? Next generation science standards, ISTE standards-S National Research Council.
3. Why do we use infographics? Require analysis an interpretation, format aids comprehension, visual learners.
4. What are tools for creating the infographics? Ease.ly, Visua.ly, Infogr.am, Piktochart.
5. What goes into infographics, what to consider when using or creating?
6. How do we evaluate infographics?

IV. EVALUATING APPROACH

Anyone can create an infographic and put it on the web. Not all infographics are good or accurate; it is like when we want to validate or evaluate a website. So we should validate an infographic carefully before using it. The downside of infographic is that data can be skewed and/or have a margin of error which would make the data irrelevant. Data is constantly changing on a daily basis, so the information presented could be outdated [1]. Although data visualizations and infographics can be created improperly; putting in too much information (or not enough), using inappropriate types for the information

provided, and unsuitable charts or graphs and other failures are common. To avoid sharing some poorly designed infographics, the audience needs evaluation and checking-questions. The things that we have to consider for infographics in general are: audience, evaluation, purpose, design and bias, readability, interactivity, and social sharing. Usually in order to evaluate how well an infographic designs and creates, there are some of discussion questions [10] to guide how creator thinks and discusses as follows:

Questions about the information:

- ✓ What is useful about putting information in this format?
- ✓ What different pieces of information are included on this poster?
- ✓ What information was included in this poster that allows non-science people to understand the content?
- ✓ Write an abstract about what it is about (two or three sentences that highlights the purpose of the infographic.)
- ✓ What is the infographic about? What story is being told?
- ✓ Does it have a clear and meaningful title? What kind of headlines, intro copy, and labels could it include to make it meaningful for a broad audience?
- ✓ Does it tell a story? What are the most important or surprising points in the data?
- ✓ What do the data mean?
- ✓ Are there sources for the data? Visit the sources? Are they valid websites/sources?
- ✓ Does it tell a story? What are the most important or surprising points in the data?
- ✓ Could we go beyond what is currently presented?
- ✓ Can we provide a better context for the data?
- ✓ Are there spelling or grammar errors? (if there are errors, chances are there are errors in the data)

Questions about aesthetics of text, photo, object, and color:

- ✓ How are colors used differently in each one? Are some colors more powerful than others?
- ✓ How are objects displayed on each one? Do sizes of the objects matter in presentation the proposed information?
- ✓ How could colors, sizes, and kinds of objects be used to give the wrong impression about people away from the data?
- ✓ Can we emphasize them by some means?
- ✓ How do the words support or distract from the message?
- ✓ How are colors used differently in each one? Are some colors more powerful than others?
- ✓ What other variables should be gathered/analyzed if we want to give an accurate portrait of the topic the graphic covers?

Questions about the overall assessment:

- ✓ Make notes about what you notice and like/don't like about the infographics.
- ✓ Is it legible? Can you read it and make sense of it?
- ✓ Can you sum up the point or message in two sentences or less?
- ✓ Who is the author? Is there any credit or information to identify the author as reputable?
- ✓ Is this infographic really “functional” in the sense of facilitating basic, predictable tasks (comparing, relating variables, etc.)?

Questions about the charts or graphs types:

- ✓ Why would this be better than just showing the formulas or using just a bar graph?
- ✓ Color and graphics? Are they legible and easy to read?

Lynda.com mentioned in the tutorial video about the five attributes of great infographic, they are contrast, hierarchy, accuracy, relevance, and truth [19]. University of Mary Washington, infographics blog presented the idea of

characteristics of an effective infographic [20] in [1] show the summaries four main categories were identified: Usefulness, Legibility, Design and Aesthetics.

Usefulness: Easy to understand, Clear purpose, Reliable data (sources cited), Informative – viewer learns something.

Legibility: Easy to read, Color scheme should not hinder ability to read, Graphs/diagrams labeled appropriately, Font choice, size and color used to make legible.

Design: Graphics should reflect purpose and audience, Graphics are good quality, not distracting and consistent, Space used effectively (no excess clutter), Appropriate use contrast and color.

Aesthetics: Easy to follow, overall design facilitates understanding, Hierarchy/organization of data.

Also we can create those questions as the questionnaire or form filled. The example of infographic evaluation form is shown in Figure 14.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
	children Strongly Disagree	children Disagree	children Neutral	children Agree	children Strongly Agree
Other (please specify)					
8. Which chapter/section did you enjoy most?					
<input type="radio"/>	chapter 1				
<input type="radio"/>	chapter 2				
<input type="radio"/>	chapter 3				
<input type="radio"/>	chapter 4				
<input type="radio"/>	chapter 5				
<input type="radio"/>	chapter 6				
9. Why? (multiple choices allowed)					
<input type="checkbox"/>	It was more amusing				
<input type="checkbox"/>	I preferred the information				
<input type="checkbox"/>	I preferred the way the story was visualised				
<input type="checkbox"/>	I preferred the characters				
Other (please specify)					
10. What is your opinion on the way the characters are visualised?					

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
I was able to recognise the characters	<input type="radio"/> I was able to recognise the characters Strongly Disagree	<input type="radio"/> I was able to recognise the characters Disagree	<input type="radio"/> I was able to recognise the characters Neutral	<input type="radio"/> I was able to recognise the characters Agree	<input type="radio"/> I was able to recognise the characters Strongly Agree
I could relate to the characters	<input type="radio"/> I could relate to the characters Strongly Disagree	<input type="radio"/> I could relate to the characters Disagree	<input type="radio"/> I could relate to the characters Neutral	<input type="radio"/> I could relate to the characters Agree	<input type="radio"/> I could relate to the characters Strongly Agree
The characters are expressive	<input type="radio"/> The characters are expressive Strongly Disagree	<input type="radio"/> The characters are expressive Disagree	<input type="radio"/> The characters are expressive Neutral	<input type="radio"/> The characters are expressive Agree	<input type="radio"/> The characters are expressive Strongly Agree
The characters made me smile	<input type="radio"/> The characters made me smile Strongly Disagree	<input type="radio"/> The characters made me smile Disagree	<input type="radio"/> The characters made me smile Neutral	<input type="radio"/> The characters made me smile Agree	<input type="radio"/> The characters made me smile Strongly Agree
Other (please specify)					

Figure 14. The example of evaluation form¹² for Alice in wonderland infographic

¹² <http://www.surveymonkey.com/r/?sm=zFaXDdelKIWwOD5cHeTAIQ%3d%3d>

IV. CONCLUSION AND RECOMMENDATION

Infographics present information in a limited space and an artistic format. They are able to quickly pass on facts and keep the readers reading it. They give important data/information and are enjoyable to read and understand. Infographics become popular in web 2.0, because they are data in graphic illustrative form which makes it easier for readers to look at and digest [17] in [18]. Infographics show how the advertising of data and information can be ordered/arranged and offered to integrate the summarize ideas [21]. For the website, with the advanced technology, infographic can be integrated with multimedia concept by adding together the sound and motion. However, as mentioned in [16] and [12], infographics are not a substitute when we do not have real information/facts. It means that ethical issue is very important, always use the actual data; they should be done without estimating at or making up data to add the missing information. Animated interactive infographics will happen to common by using HTML5 in websites; it makes infographics much more interesting.

Considering the infographic as a tool for helping human extend capacities beyond the brain limitation, physical tools like tangible things can help human life easier, such as Notebook [22], Mobile Phone, etc. Non-physical tools (or sets of tools and practices), for instance statistical data and the scientific method, developed for helping the audience look beyond what would normally see, and to overcome biases and some negative habits of mind. The same is true for great visual displays of information.

When constructing an infographic, one must consider how best to get across the information and the purpose. Elements of design factor into this; the visuals need to support the information and should not detract from it. The purpose of the infographic informs how information is presented (i.e. statistically, linearly, and procedurally) as well as the images used. When people interact with infographics in two primary ways which are as consumers and as creators of learning (or cognitive processes) measure when human is ready to move from a consumer to a creator. Humans must be able to remember and understand information before they are able to analyze or evaluate and finally synthesize or create it. Additionally, there are certain skills necessary to create an infographic, both technical and literacy-based. We can see from all the questions for infographic evaluation that there are many questions about what information and data should be used and they have to be correct. The presentation should be only beautiful and attractive, but that is a secondary component of their quality. They are, above all, accurate [10]. Every single detail is double-checked and reviewed by experts; every line, patch, and shade of color has a meaning and a function. So the evaluation is matter because it can help creating the good infographic and make it easier for the audience/user/people to understand the main idea of the infographic.

REFERENCES

- [1] Thatcher, B. (2012), An Overview of Infographics, Webinar. Illinois Central College Teaching & Learning Center. Retrieved from www.slideshare.net/iccitic2
- [2] Siricharoen, W. V. (2013). Infographics: The new communication tools in digital age. International Conference on E-Technologies and Business on the Web, Bangkok, Thailand, p. 169-174. Retrieved from <http://sdiwc.net/digital-library/infographics-the-new-communication-tools-in-digital-age>
- [3] www.mediakar.org. (2013). How to evaluate infographics, Data journalism: Visualization: Digital storytelling. Retrieved from <http://mediakar.org/2013/03/22/how-to-evaluate-infographics/>
- [4] Blueprint, 2013, Tools for Creating High-Quality Infographics Your Own, Retrieved from http://www.slideshare.net/blurbpoint/tools-for-creating-high-quality-info-graphics-your-own?next_slideshow=1
- [5] Spry, K. C. (2012). An infographical approach to designing the problem list, IHI '12: Proceedings of the 2nd ACM SIGHT International Health Informatics Symposium.
- [6] Ashton, D. (2013). The 8 Types of Infographic. Retrieved from <http://neomam.com/infographics/the-8-types-of-infographic/>
- [7] www.Branded4good.com. (2012). 8 Types of Infographics and How Nonprofits Can Use Them. Retrieved from <http://branded4good.com/blog/8-types-infographics-nonprofits/>
- [8] Nitsche, L. (2013) Infographics: Analyze, Evaluate and Create. Retrieved from <http://www.slideshare.net/lnitsche/infographics-analyze-evaluate-and-create-16443121>
- [9] www.webdesignerdepot.com. (2009). 30 Superb Examples of Infographic Maps Design, Inspiration, Web Design. Retrieved from <http://www.webdesignerdepot.com/2009/10/30-superb-examples-of-infographic-maps/>
- [10] Starr, K. (2014). Infographics. Retrieved from <https://classes.lt.untd.edu/...2014/.../Starr%20expanded%20storyboard.doc>
- [11] Pavlus, J. (2013). The Future Of UX Design: Tiny, Humanizing Details. Retrieved from <http://www.fastcodesign.com/1672922/the-future-of-ux-design-tiny-humanizing-details> [2013-07-02]
- [12] Kocher, J. (2012). The SEO Benefit of Infographics. Retrieved from <http://www.practicalecommerce.com/articles/3836-The-SEO-Benefit-of-Infographics>
- [13] Golombisky, K., Hagen, R. (2010). Chapter 11 – The Scoop on Infographics: Maximum Information in Minimum Space, White Space is Not The Enemy A Beginner's Guide to Communicating Visually through Graphic. Web and Multimedia Design, p. 153–166, 2010.
- [14] Beck, H. (1951). Paris Metro by Harry Beck. Retrieved from <http://tinyurl.com/bnh7qe>
- [15] www.visual.ly.com, (2014). Retrieved from <http://visual.ly/learn/what-are-interactive-infographics>
- [16] Siricharoen W. V.(2013). Infographics: An Approach of Innovative Communication Tool for E-Entrepreneurship Marketing. IJEEI 4(2). p. 54-71 (2013).
- [17] Byrne, R. (2011). Picture this, School Library Journal 57(6), p. 15
- [18] Moorefield-Lang, H. (2011). Infographics: Information Gets Visual. Information Searcher, 19(3), p. 15-16.
- [19] www.lynda.com. (2013). Illustrator tutorial: The five keys to a great infographic. Retrieved from <http://www.youtube.com/watch?v=UQwEEoqLrk>
- [20] www.infographics2011.umwblogs.org. (2011). Rubric for Effective Infographics, UMW FSEM Infographics. Retrieved from <http://infographics2011.umwblogs.org/2011/11/16/rubric-for-effective-infographics/>
- [21] Joss, M. (2013). Book Reviews: Infographics: the Power of Visual Story Telling”, The Seybold Report, 13(3), p. 8-12.
- [22] Cairo, A.(2012). Infographics and Visualizations as Tools For the Mind. Retrieved from <http://blog.visual.ly/infographics-and-visualizations-as-tools-for-the-mind/>
- [23] Krum, R. (2014). Effective Communication with Data Visualization and Design, John Wiley & Son, Inc. Indianapolis, Indiana. p.8.

Noise Reduction Algorithm Based on Complex Wavelet Transform of Digital Gamma Ray Spectroscopy

Mohamed S. El_Tokhy

Electrical Engineering Department, College of Engineering,
Aljouf University,
Aljouf, KSA
engtokhy@gmail.com

Imbaby I. Mahmoud

Engineering Department, NRC,
Atomic Energy Authority,
Inshas, Cairo, Egypt

Abstract—This paper investigates the use of complex wavelets in gamma ray spectroscopy signals. In this paper an algorithm for noise elimination of the detected gamma ray spectroscopy signals is studied. This algorithm is based on the complex wavelet transform. Reconstruction of the original detected signal is obtained by applying the inverse complex wavelet transform to the transformed complex wavelet transform signal. Five different cases are studied with different five levels of the complex wavelet transform. Consequently, comparisons between these levels are considered in terms of maximum number of peak heights, execution time, and peak signal to noise ratio (PSNR). Moreover, comparison between different signal reconstruction with respect to different complex wavelet transform levels, size of the transformed signal in each level, and number of coefficient in each subband for certain level. One of the main advantages of this algorithm that discussed in the previous literature is that its filters do not have serious distributed bumps in the wrong side of the power spectrum and, simultaneously, they do not introduce any redundancy to the original signal. The obtained result confirms the high accuracy of the considered algorithm over traditional algorithms for both noise elimination and signal reconstruction.

Keywords—Complex Wavelet Transform; Peak Signal-to-Noise Ratio; Linear Filters

I. INTRODUCTION

THE Scintillation detection experiments indicate that the electronic noise in InI photodetectors was the dominant source of resolution broadening [1]. Hence, the electronic noise behavior of the InI detectors was investigated to determine the magnitude of various noise components in the detectors. These existing noise models are used to analyze the electronic noise in InI detectors. The electronic noise is expected to arise from several sources in InI detectors [1]. These sources are the parallel thermal noise due to the detector leakage current (also commonly referred to as shot noise).

Secondly, the series thermal noise that generated in the channel of the input JFET of the pre-amplifier. Finally, the 1/f noise has been obtained from the detector pre-amplifier assembly. Since the PMT anode signal is very noisy and timing features highly depend on the signal at specific times, a de-noising algorithm is required.

There exist different digital de-noising and smoothing methods (e. g. moving average filters) depending on the application [2]. A de-noising algorithms based on the Wavelet Transform (WT) is implemented to reduce the effect of noise introduced by the noisy analog channel and by the photomultiplier tube as in [2]. In this application, linear smoothing filters are not appropriate because the signal contains a sharp portion associated with the interaction in the first layer (fast component).

Wavelet de-noising which is a non-linear filtering operation analyzes the signal at different time resolution levels and then removes the noise components by thresholding signal components in one or more levels. Depending on the application, the level and threshold should be modified to remove the noise while keeping the important high-frequency components of the signal [2].

Complex wavelet transforms has significant advantages over real wavelet transform for certain signal processing problem [3]. Complex wavelet transforms, in which the real

and imaginary parts of the transform coefficients are an approximate Hilbert-transform pair, offer three significant advantages over real wavelet transforms: shift invariance, directionality, and explicit phase information. These properties enable efficient statistical models for the coefficients that are also geometrically meaningful [4]. Complex wavelets have not been used widely in signal processing due to the difficulty in designing complex filters which satisfy a perfect reconstruction property [5]. To overcome this Kingsbury [6] proposed a dual-tree implementation of the CWT (DT CWT) which uses two trees of real filters to generate the real and imaginary parts of the wavelet coefficients separately. Even though the outputs of each tree are downsampled by summing the outputs of the two trees during reconstruction, the aliased components of the signal is suppressed and achieved approximate shift invariance [5]. The complex wavelets are first used to perform analysis of the signals [5]. We describe how to extract features to characterize textured signals and test this characterization by resynthesizing textures with matching features. The term de-noising is usually referred to removing the white Gaussian noise or thermal noise which is added to the signal. In our application, this type of noise is mostly introduced by the noisy analog read-out system. In our application, a 5-level de-noising algorithm based on complex wavelet transform functions was used. Also, the rescaling in wavelet decomposition is performed using level-dependent estimation of the noise level. This paper is organized as follows: Section 2 presents the spectroscopy system. The more interesting characteristics of the studied complex wavelet transform are represented in Section 3. Results and discussion are summarized in Section 4 and we terminate our study by a briefly conclusions that we noted from our obtained results.

II. SYSTEM CONFIGURATION

In this system, the components of the system for evaluation of noise elimination using complex wavelet transform algorithms are described. Contains the following elements; ¹³⁷Cs point source, scintillation detector, amplifier, digital system and connection to a desktop personal computer (PC). An 1.5 inches x 7.5 inches NaI(Tl) scintillation detector is used to detect the radiation signal from Cs137 point source. This detector is connected to amplifier through coaxial cable which in turn connected to the PC. MATLAB environment is used to perform noise elimination using complex wavelet transform evaluation.

In this paper, an algorithm for noise elimination using complex wavelet transform is studied on digital gamma ray spectroscopy signals. This algorithm is proposed for multidimensional signal processing in [7]. Block diagram showing the algorithm of noise elimination evaluation using

the complex wavelet transform is illustrated in Fig. 1. Moreover, different wavelet transform levels are considered.

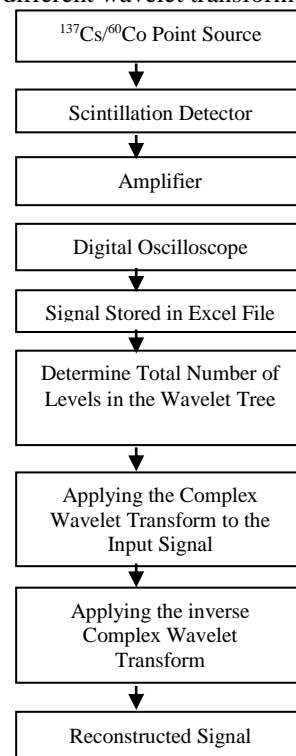


Fig. 1 Block diagram model of the complex wavelet transform algorithm for noise elimination of gamma ray spectroscopy signal

III. THE COMPLEX WAVELET TRANSFORM

A. Linear Phase Filters

One of the most important properties of the filters which can be applied to 3-band filter banks is linearity of the phase. The phase of the signals which are filtered using the linear phase filters is not perturbed, which means all frequency components of the signal are shifted equally. The filter $h(t)$, which has the linear phase property, satisfies the following equation[7]

$$h(n) = e^{i\theta} h^*(N-1-n), 0 \leq n < N-1 \quad (1)$$

where N and θ is the length of the filter and an arbitrary variable between zero and 2π . The following four special kinds of the linear phase filters are used:

1. Real filter with $\theta = 0, 2\pi$ and N to be odd,
3. Complex filter with $\theta = \pi/2, 3\pi/2$

B. Filter Bank Characteristics

Filter banks are widely used in digital signal processing, often integrated in a multirate scheme, to reduce the implementation cost and to improve algorithmic performance [8]. The ideal normalized power spectrum of the filters in the filter bank that satisfies all desirable characteristics, such as having the Hilbert-pairs wavelet

filters and introducing no bumps on the wrong side of the power spectrum. This filter bank introduces no redundancy because one of the wavelet filters is the complex conjugate of the other, so the resulting coefficients of one filter is the complex conjugate of the other filter, and we can discard them in further analysis [7].

In the ideal case, without considering this low-pass filter, the sampling frequency can be computed using Nyquist's theorem. In the non-ideal case, an anti-aliasing filter must be applied before the sampling part. In this case, the sampling frequency can be calculated using the following equation [7]:

$$f \geq f_n + 2B_{tr} \quad (2)$$

where f_n is the Nyquist frequency and B_{tr} is the transient band of the anti-aliasing filter. When using a discrete time low-pass filter after a continuous-to-discrete converter, we can compute another constraint for avoiding distortion. The sampling rate must be computed in such a way that the normalized bandwidth of the analog signal be equal to the pass-band of the low-pass filter. Therefore, the following constraint is computed [7]:

$$f_s \geq \frac{f_n}{2B_{low}} \quad (3)$$

where B_{low} is the normalized pass-band of the low-pass filter. In this case, the amount of sampling frequency increase with respect to the common sampling is [7]:

$$R = \frac{f_n / 2B_{low}}{f_n + 2B_{tr}} \quad (4)$$

In addition, natural signals decay very fast with frequency increase; therefore if we deviate a little from this sampling rate constraint, the amount of distortion introduces into signal is not very much. We will see this in the second part of the simulation results.

C. Filter Bank Design Procedure

An orthogonal filter bank for digital gamma ray spectroscopy which consists of one real filter $h_0(n)$ and two complex conjugate filters $h_c(n)$ and $h_c^*(n)$ is studied. It is considered [7]

$$h_c(n) = \sqrt{\frac{1}{2}} \times (h_1(n) + ih_2(n)) \quad (5)$$

then for having a complete orthogonal transform, $h_0(n)$, $h_1(n)$ and $h_2(n)$ must satisfy the shift orthogonal condition expressed in the next equation.

In order to have a complete orthogonal transform, the scaling and the wavelet functions must satisfy the shift orthogonal condition.

The design cost function is given by [7]

$$\Phi = \left(\sum_{m=0}^2 \sum_l \sum_{i=0}^2 \{h_i^T Q_l h_m - \delta(l) \delta(m-i)\}^2 + \alpha \left\{ \begin{aligned} &h_0^T M_0(\omega) h_0 + h_1^T M_{re}(\omega_2) h_1 + \\ &h_2^T M_{re}(\omega_2) h_2 + 2h_2^T M_{im}(\omega_2) h_1 \end{aligned} \right\} \right) \quad (6)$$

where α is the weighting coefficient and the stopband power is calculated at those frequencies ω_1 and ω_2 that maximize it. For this purpose, the stopband is sampled uniformly and the stopband power is evaluated at these sampled frequencies. A gradient descend algorithm for minimizing the cost function is used. At every iteration the weighting coefficient α should decrease such that the optimized filters satisfy the shift orthogonal condition. This iteration will terminate when the cost of the shift orthogonal condition expressed in the first line of Equation 6 becomes sufficiently small ($\approx 10^{-6}$).

IV. RESULTS AND DISCUSSION

We apply the low-pass filter before the analysis and after the synthesis filter banks, therefore we deviate from the perfect reconstruction condition. To observe the amount of distortion introduced into the signal, we apply different levels of the filter bank with the low-pass filter to the acquired signal. The original detected signal is depicted in Fig. 2.

The impulse response of the filters at levels one, two, three, four, and five at different shifts and the resulting absolute value of the wavelet coefficients at these levels are shown in Figs. 3-7, respectively. As illustrated in these figures, these filters are highly oriented in 00, ± 450 , 900, and 1350 and the real and complex parts of the complex filter constitute Gabor-like filters as in [7]. As illustrated, the filters even at the first level are oriented. It is interesting to test the shift-invariance performance. It is obvious that the studied filter bank has an excellent shift-invariance property which is comparable to the dual-tree complex wavelet transform. For the studied filter bank with the filters of length 37, R_a is calculated and is shown in Table 3. As an illustrative example, we construct the signal of one, two, three, four, and five levels and the reconstructed signals are shown in Figs. 6-12. One can see that transformed signals are shift-invariant and free of aliasing.

Comparison between the different wavelet levels are depicted in Table 1. From this table, the number of counted peaks decreases with the wavelet level. Also, the number of counted peaks of the original signal is equal to the number counts of level five complex wavelet transform. However, the execution time increases with the level. Also, comparison between different five levels is illustrated in terms of PSNR. From the theoretical results, the PSNR of the reconstructed signals are very high. Therefore, the

underlined filter bank is considered as a nearly perfect reconstruction filter bank on natural signals. Moreover, comparison between different signal reconstruction in terms of different complex wavelet transform levels, size of the input signal, size of the transformed signal in each level, number of coefficient in each subband for certain level, and total length of the input signal is depicted in Table 2. Also, the normalized amplitude of level 1 complex wavelet transform is depicted in Fig. 13.

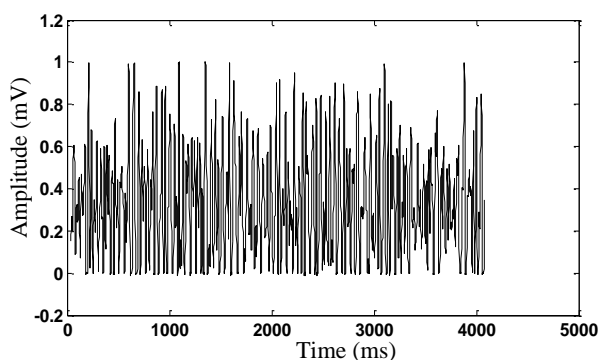


Fig. 2 Original detected signal from the scintillation detector

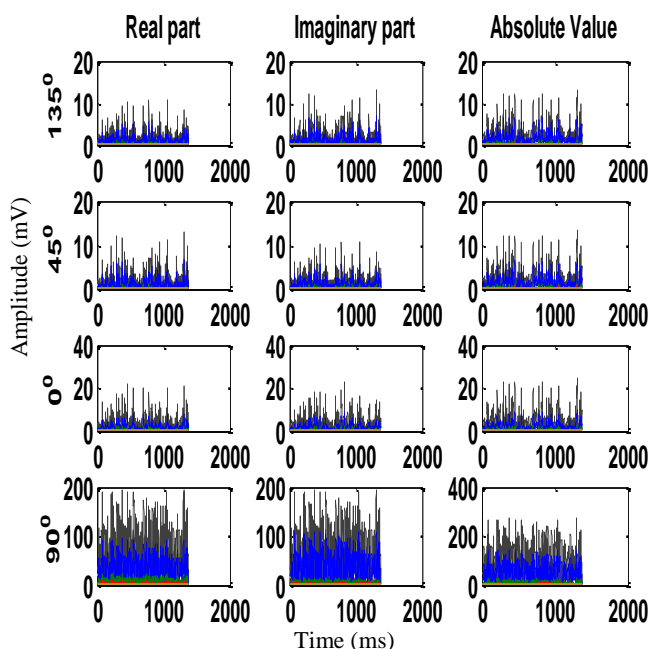


Fig. 3 The impulse response of the filters at level one at different shifts and the resulting absolute value of the wavelet coefficients

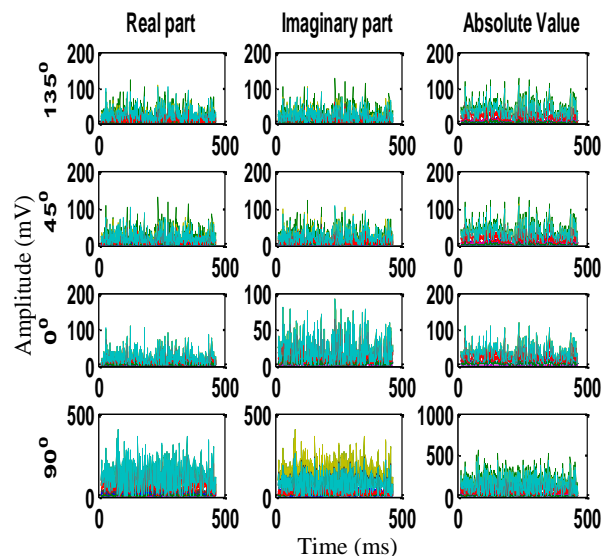


Fig. 4 The impulse response of the filters at level two at different shifts and the resulting absolute value of the wavelet coefficients

TABLE I
 COMPARISON BETWEEN DIFFERENT COMPLEX WAVELET TRANSFORM LEVELS

	Number of Peaks	Execution Time (s)
Level 1	207	4.015
Level 2	188	4.516
Level 3	172	4.829
Level 4	147	5.266
Level 5	143	9.25

TABLE 2
 COMPARISON BETWEEN DIFFERENT LEVELS IN TERMS OF DIFFERENT STRUCTURE PARAMETERS

	Subbands	Coefficient Number	Total length
Level 1	[4082 1]	17849	160641
Level 2	[2x2 double]	[17849 5115]	188827
Level 3	[3x2 double]	[17849 5115 1782]	199750
Level 4	[4x2 double]	[17849 5115 1782 671]	204007
Level 5	[5x2 double]	[17849 5115 1782 671 297]	206009

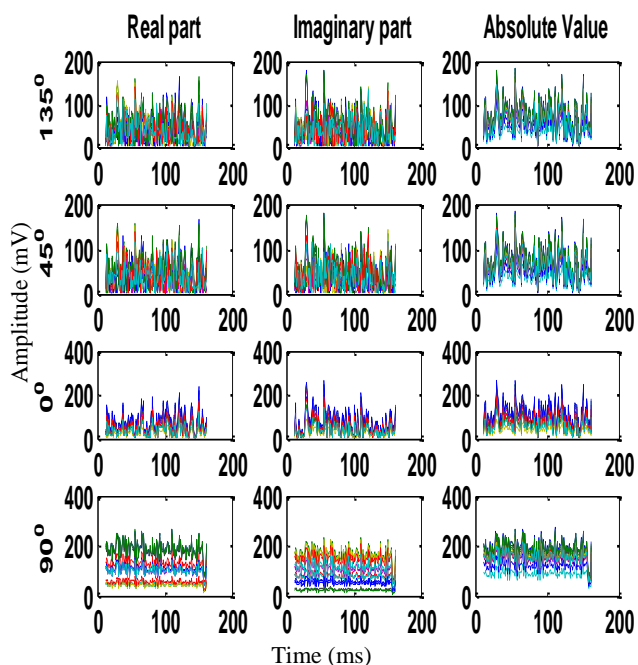


Fig. 5 The impulse response of the filters at level three at different shifts and the resulting absolute value of the wavelet coefficients

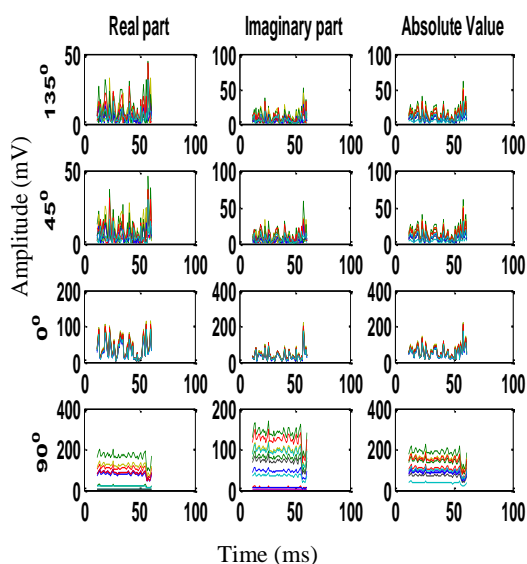


Fig. 6 The impulse response of the filters at level four at different shifts and the resulting absolute value of the wavelet coefficients

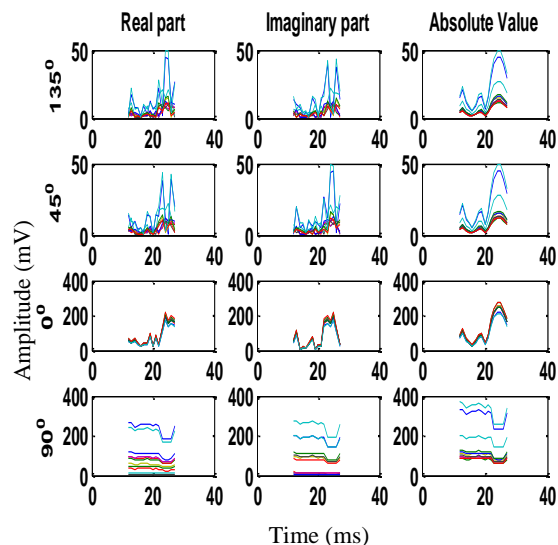


Fig. 7 The impulse response of the filters at level five at different shifts and the resulting absolute value of the wavelet coefficients

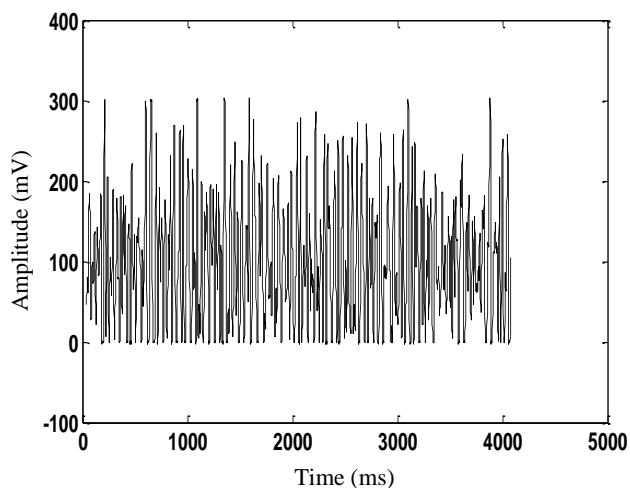


Fig. 8 Reconstructed signal using first level of the algorithm.

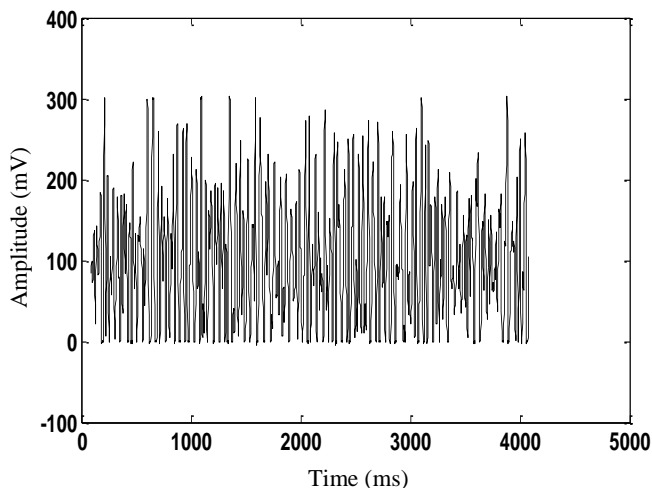


Fig. 9 Reconstructed signal using second levels of the algorithm.

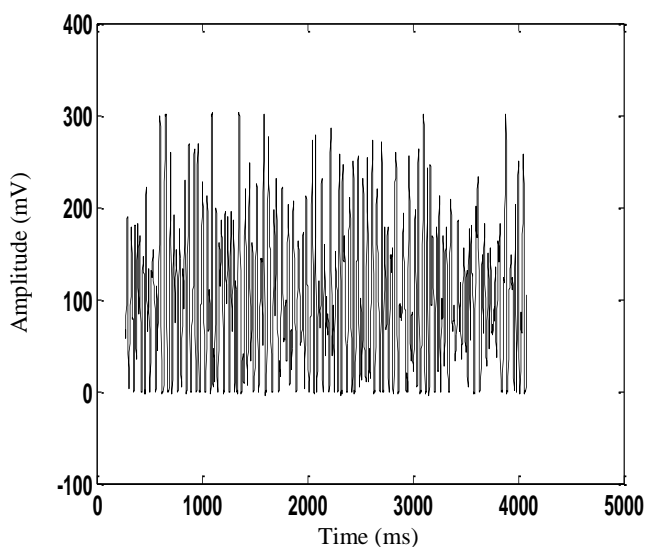


Fig. 10 Reconstructed signal using three levels of the algorithm.

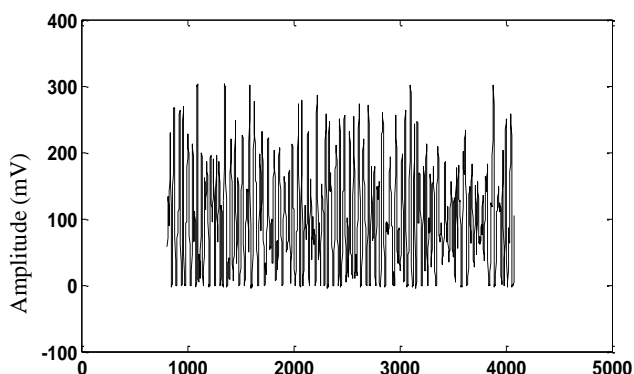


Fig. 11 Reconstructed signal using four levels of the algorithm.

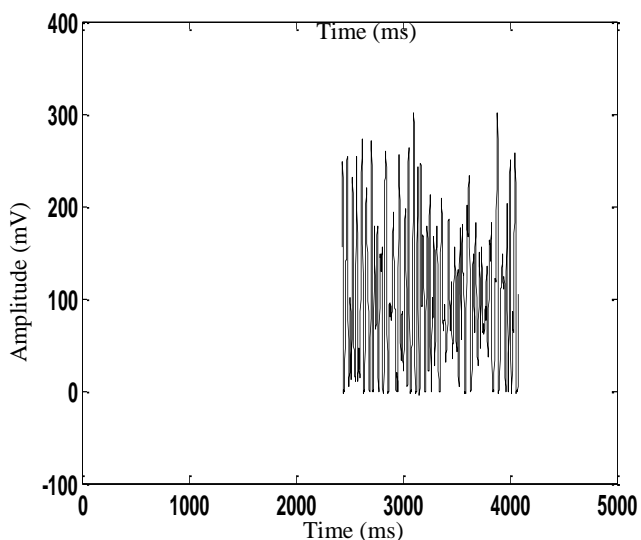


Fig. 12 Reconstructed signal using five levels of the algorithm.

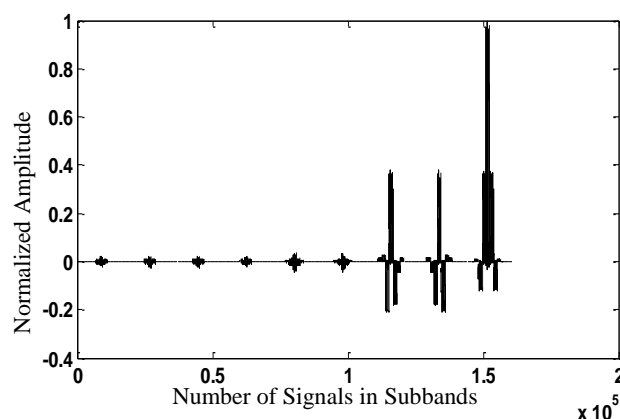


Fig. 13 Normalized amplitude of the output signal.

CONCLUSION

This paper focuses on noise elimination due to both electronic system readout and coaxial cables between scintillation detector and amplifier. An algorithm based on complex wavelet transforms is studied to do this function. The input signal is transformed using the complex wavelet transform. Then, the inverse complex wavelet transform is applied to the transformed signal. Comparison between different complex wavelet transform levels is considered. This comparison is based on both number of counted peaks, execution time and PSNR. This filter bank has linear phase filters. The resulting filters at different levels do not produce serious bumps on the wrong side of the frequency axis.

REFERENCES

- [1] K. S. Shah, P. Bennett, L. P. Moy, M. M. Misra, W. W. Moses, "Characterization of Indium Iodide Detectors for scintillation studies", Nuclear Instruments and Methods in Physics Research A, Vol. 380, pp. 215-219, 1996.
- [2] Siavash Yousefi and Luca Lucchese, "Digital Pulse Shape Discrimination in Triple-Layer Phoswich Detectors Using Fuzzy Logic", IEEE Transactions on Nuclear Science, Vol. 55, No. 5, October 2008.
- [3] H. N. Abdullah, "SAR image denoising based on dual-tree complex wavelet transform", Journal of Engineering and Applied Science, Vol. 3, No. 7, pp. 587-590, 2008.
- [4] Felix C. A. Fernandes, Michael B. Wakin and Richard G. Baraniuk, "Non-redundant, linear-phase, semi-orthogonal, directional complex wavelets", ICASSP, IEEE, 2004, pp. II - 953- II - 956.
- [5] Musoko Victor, Prochazka ales, "Complex wavelet transform in signal and image analysis", 14th International Scientific - Technical conference on Process Control, Czech Republic, June 2004.
- [6] N. G. Kingsbury, "Complex wavelets for shift invariant analysis and filtering of signals", Journal of Applied and Computational Harmonic Analysis, Vol 10, No 3, pp. 234-253, 2001.
- [7] Reshad Hosseini, Mansur Vafadust, "Almost Perfect Reconstruction Filter Bank for Non-redundant, Approximately Shift-Invariant, Complex Wavelet Transforms", Journal of Wavelet Theory and Applications, Vol. 2, No. 1, pp. 1-14, 2008.
- [8] Koen. Eneman, and Marc. Moonen, "DFT modulated filter bank design for oversampled subband systems", Signal Processing Journal, Vol. 81 No.9, pp.1947-1973, 2001.

Designing Children's Encyclopedia (3D Dinosaur)

Via Augmented Reality Marker-Based Interaction

Anita Mohd Yasin

Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA(UiTM)
Shah Alam, Malaysia
anitamy@tmsk.uitm.edu.my

Zeti Darleena Eri

Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA(UiTM)
Terengganu, Malaysia
zeti415@tganu.uitm.edu.my

Mohd Ali Mohd Isa

Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA(UiTM)
Shah Alam, Malaysia
ali@tmsk.uitm.edu.my

Nor Adora Endut

Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA(UiTM)
Shah Alam, Malaysia
adora@tmsk.uitm.edu.my

Abstract—The purpose of this project is to design a new form of interface for children's learning by applying the augmented reality (AR) technology. Our main idea on this project is to explore the method of interaction that most people have yet to know about. The existing methods of user interface have become quite common to the users. Hence, we provide the new approach of interaction called AR Marker-based interaction in solving difficulties in attracting and cultivating children's interest in learning. This study is implemented in a children's 3D Dinosaur Encyclopedia. The motivation for this study is due to the lack of interaction and interest in the common teaching methods. Additionally, our study aims at providing new ways to make learning fun for children.

Keywords— *augmented reality, marker-based interaction, children*

I. INTRODUCTION

Previously, augmented reality (AR) is known as a part of mixed reality (MR) in the field of virtual reality (VR). But AR and MR have been seen as the same terms of the technology use. AR is reaching its significance as the new medium in the evolution of MR [1]. However, AR and MR have increasingly been seen as the same in terms of the technology being used. MR is reaching its significance as the new medium in the evolution of VR? The usage of AR has been seen in various fields including architecture, advertising and navigation

systems application. Recently, there has been a marked increase in the use of MR in outdoor environments [2].

The need to develop augmented reality applications that assist in learning is because of unexploited dynamic interactive visual imagery [3]. Using augmented reality applications as learning assistants can exploit dynamic visualization. Furthermore, [3] stated that the use of augmented reality will make learning experience more enhanced and users will have better interaction.

In applications for children, AR has been used to help in visualization. Among others are those which have been developed by Engine Design Puzzle Material Master and Materials Mastermind. Additionally, there are also storybooks in augmented reality that use multimedia elements to represent the story for kids [3]. These applications provide the visualization and interaction that help children to understand the contents. In this project, the development of the AR application is focusing on assisting the children to visualize effectively the content that originally comes in printed format. In a study done by [4], 3D models are overlaid over printed book pages or 2D images which have been captured using personal computer camera. This approach will be used in our study to display the types of dinosaurs and to provide a method for interaction as existing AR application.

There are currently many applications available for learning about dinosaurs in AR form for children. However, they still lack the features to provide means for independent learning. Although there is some form of interaction, its only purpose is to control the view. On top of that the interaction that these applications promote do not convey the information contained in the encyclopedia.

2D illustrations found in printed or book pages may be less effective to convey the context of some objects and it may lead to an incorrect interpretation, particularly those related to spatial representations [5]. 3D visualization contains realistic and detailed objects; and may change to provide first-person perspective as compared to 2D visualization which focuses on flat spatial representation [6]. The 3D graphics can affect the effectiveness and efficiency of how the audience views the graphics [7] and demonstrated better viewed as shown by many studies [8]. In terms of interaction, with recent development of 3D data acquisition there is a possibility of a more efficient manner to reproduce real objects into virtual images and enable users to better manipulate and interact with these images [9].

Augmented reality development aims to provide a new experience to the community to see digitally enhanced objects in a real environment [10]. Since augmented reality is relatively a new form of technology, it can be adapted for learning assistants in the real environment [3]. This technology can be used to better engage students in the teaching materials more effectively. Furthermore, augmented reality enhances children's understanding of dinosaur encyclopedia based on the book through interactivity. It will benefit learners through the capabilities for the users to control, manipulate and share information subject.

II. RELATED WORK

A. Augmented Reality (AR)

Augmented Reality is a part of virtual reality or virtual environment as its alternative name. Virtual Environment totally brings the user into an artificial world and user is unable to see the real environment. In Augmented Reality, users are able to see the real environment with virtual objects placed on the real environment. In augmented reality the real and virtual objects are superimposed in the same space. Augmented reality can be perceived as the "mediation" of virtual reality and

telepresence which is immersive and realistic respectively [11]. Augmented reality was as subset of Mixed reality since they have the similarities of the way they used technologies in which mixed reality technologies will be used as the new medium [1]. Augmented reality was a subset to the virtual reality in which virtual reality fully utilizes the virtual environment seems like the real environment while in augmented reality, user can see virtual objects in their real environment [12]. According to [4], there are also various forms of media that use augmented reality technologies using the printed pages as the markers.

B. Interaction

The interaction mode for augmented reality is named Exploring [13]. The exploring interaction allows user to move through a virtual or real environment that may come from virtual and augmented reality systems. In the physical environment, the system will be embedded with sensing technologies in which the sensing device will detect the presence of physical object, and respond by executing the digital events.

C. 3D Graphics Visualization Enhanced with Multimedia Elements in Augmented Reality

3D visualization or graphics is a presentation and the control of objects in the spatial presence in a computer in which the application program for designing 3D objects give the properties such as height, width and length and the images can be rotated and scaled. There has been an increase in the use and influence of the 3D visualization digital presentation and technology in people's life [14]. Furthermore, the latest technology improvements in 3D data gathering and interaction allow for efficient generation and manipulation of complex real life objects [9]. There are several augmented reality applications that used 3D visualization. In the medical field, the surgeon uses the augmented reality application to visualize the anatomy of the affected area by designing the 3D graphic from the various views and parts. In the engineering field, the prototype will be designed in 3D and displayed and imaged in the client's conference room. Another example will be for the education: Construct 3D in which it is for math learning and geometry in 3D. Although most augmented reality application displaying 3D visualization, the augmentation can be represented in both 2D and 3D graphics [15].

D. 3D Graphics for Children

1) 3D Graphics Perception with Children

Children have their own perception toward seeing the 3D graphic visualization. Children often used to see 3D graphics as more fun when solving a problem although it takes time on understanding how to use the 3D representations [16]. Since the children have understood how to interact with the 3D object, it would be effective in terms of the delivery of the content. 3D scenes may have the characters and other objects that occur in 2D scene. Although the 3D scene displayed its content in a 2D view, children might feel that they are immersed in the 3D spatial environment [17].

2) Multimedia Driven for 3D Graphic in Augmented Reality

The use of 3D graphics solely is not enough to enhance children’s experience towards some multimedia content. Commonly, multimedia content such as a courseware is embedded with several elements in which text, sound, and some animation are also incorporated. As a result, the integration of these elements would make the courseware effective in delivery of their content. In augmented reality, the effectiveness of the content could be improved by enhancing it with several media such as text and sound [18].

III. RESEACRH METHOD

In our study, the following development stages have been used to develop our application. The stages involved were 3D modelling and animating, augmented reality authoring for tracking and interactivity, and lastly the completion. Below were the revised development stages with the addition of prototyping at the first stage.

Fig. 1 shows the process involved in 3D Modelling and Animation followed by the scripting for functionalities. This will be followed by the review of the AR scenes. This process will loop until the process end.

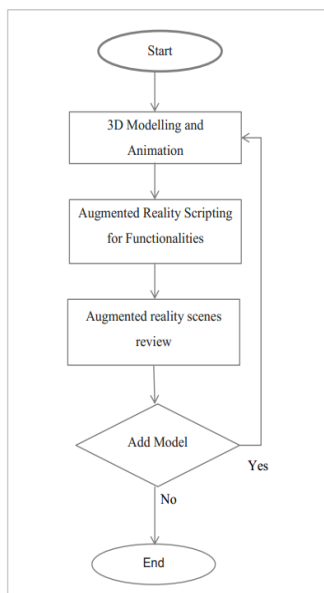


Fig. 1. The Development Flowchart

Stages towards implementation for the design of the AR are describes as follows:

A. Phase 1: Prototyping

Prototyping in the development of the AR interface had been designed in order to test what kind of interface will be produced. In this stage, the object used was the basic 3D model such as a cube, sphere and a teapot that have been made in the 3D tool which is 3D Studio Max. Fig. 2 illustrates the prototyping result.

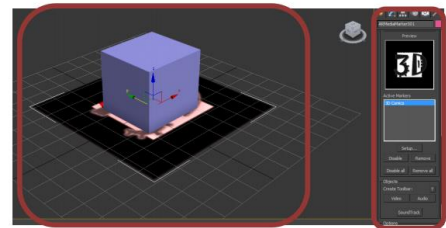


Fig. 2. Prototyping with 3D modelling tool and augmented reality plugin

The illustration above was the process of initial prototyping in which at the right highlighted section shows the option of the augmented reality plugin for 3D modelling tool. The option is used to select the provided markers that are going to be attached to the 3D model. The left highlighted section shows the modelling creation. At this section, the model was positioned above the markers.

B. Phase 2: Modelling and Animation

The dinosaur character should have been designed as the main content for the markers inside the encyclopedia. The characters are limited to ten in which each dinosaur character was designed from each of the dinosaur family. Fig. 3 and 4 illustrates the process of characters modelling.

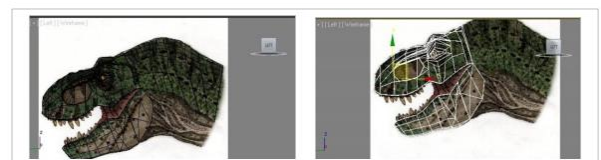


Fig. 3. Character head modeling



Fig. 4. Animation Creation

In the animation process, character bones needed to be created in order to make the animation movements realistic. The animation was controlled by adding the key frames in the modelling space. It was created repeatedly where the scenes would not stop; the character will walk around the character’s base. The modelling and animation process were applied the same way to the other characters that reflected to dinosaur facts.

C. Phase 3: Authoring

The augmented reality authoring is a process of involving those models inside the augmented reality scenes. The plugin uses XML language to create the interaction of the augmented reality. In Fig. 5, the targeted marker refers to the marker that is being used to augment characters inside the scene. A single

marker may involve several models since the targeted marker will not display only a character but the character base, character ground, character's name label, and feedback content.



Fig. 5. The involvement of character model in AR scene

IV. RESULT – COMPLETION

The final stage of the augmented reality implementation was the completion stage. Completion refers to the attachment of the character markers for the character models on the pages of each of the dinosaur's facts pages creation. For the facts, the content was taken from several websites that were suitable for children and printed on the pages. The design of the pages is constructed purely for the purposes of testing to achieve our research objectives. Below are the illustrations of the printed pages.

In Fig. 6, the dinosaur's picture was printed together with the character markers on a page. The marker was sectioned at the grey rectangle was used to make the marker visible since the pictures have various colors that may distract the marker visibilities.



Fig. 6. Attached markers on pages for character model

Figure 7 shows a page that displays the facts about the dinosaur. The marker was not attached on this page and the use

of white fonts is due to the readability factor. The printed pages explained the details of each dinosaur.

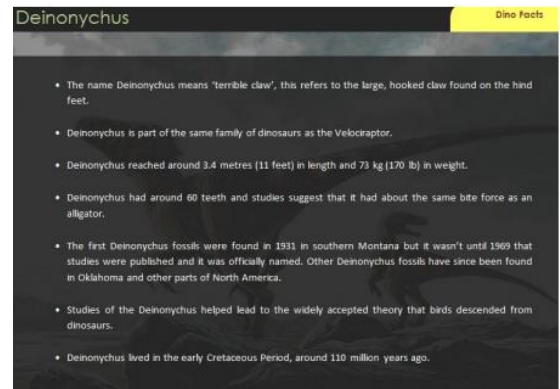


Fig. 7. The dinosaur's facts on the printed pages

Fig. 8 shows the interaction marker that was used to interact with the character model that has been stated at the design stage. The interaction marker would be printed and attached to the mounting board in order for the user to handle it easily. It was not being attached to the printed pages as it is separated from the pages. The marker is also movable to encourage the user to interact with every dinosaur character of each page.



Fig. 8. Interaction Marker Template

The interaction marker has its own categories which are carnivore marker, herbivore marker, and the size comparison to real object. In order to encourage the user on how to interact with the main characters, messages were printed on the back of the pages of the interaction marker that asks the user to bring these markers to the tracking space during the main characters AR displays.

A. Testing

1) Method 1: Testing without AR Capability

The testing was made in public environment in which 10 students were tested to use the dinosaur encyclopedia that have no augmented reality application. 10 general questions were provided for the purpose of testing the effectiveness of how the encyclopedia conveys the facts to the targeted audience. The

question was about general facts of dinosaurs. Below was the result of the testing.

Fig. 9 shows that the encyclopedia without an augmented reality in which 9 out of 10 students answered the questions correctly. Most of the questions comes from the illustration and the facts on the printed pages. There was also another question that was not from the printed pages that used to test their knowledge about dinosaurs.

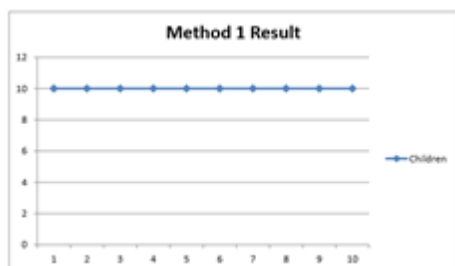


Fig. 9. Without AR capability

2) Method 2: Testing with AR Capability and no Interaction

In this method, there were 3 children that have been tested for the AR application in which the application have no interaction. The targeted audience can only view the augmented reality scenes of dinosaur animation. Furthermore, there were 11 questions provided to them in which they need to answer the questions based on the animation and the information inside the printed pages. Below is a figure that shows on how many questions that have been answered correctly a measure of effectiveness.

Fig. 10 shows the result for Method 2. The findings show that the first child answered all the 11 questions in which 6 questions out of 11 were answered correctly. The other children answered more than 6 questions correctly. From the observation, the entire 3 targeted user was able to view and control the augmented reality scenes excitedly. Although, they tend to use the augmented reality scenes compared to reading the information inside the printed pages.



Fig. 10. With AR capability and no interaction

3) Method 3: Testing AR with Interaction

This testing method is similar to the second method. The only difference is the AR application has the interaction element to convey the information at the printed pages in the form of a 3D scenery. Fig. 11 shows the improvement of the children in answering the 11 questions about the dinosaurs.

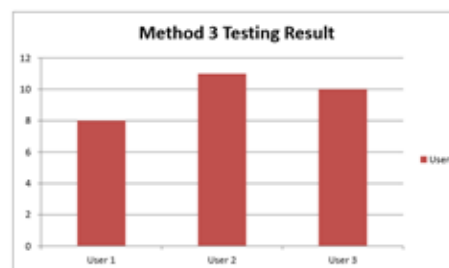


Fig. 11. AR capability with interaction

The results show that after using the AR with the interaction element involved, the first child answered 2 more questions correctly. Some users has answered all the questions correctly. This shows that the 3D visualization has helped some the children to visualize and imagine the real facts of the dinosaurs. On the other hand, only one user had no improvement because of capability of their understanding was different. The children have used the augmented reality according to their own level of interest. This means they used to seek the answer by exploring each of the printed pages and the interactions.

V. CONCLUSION

The design of AR can improve the children's experience in learning especially in supporting the visual conceptualization of various forms. It also possibly will extend and give alternative opportunities beyond our imagination in learning. Currently, the interaction that has been designed was at the minimum capability in which it can be improved to be more interactive with realistic tangible interaction that may provide different experience of interactions. Furthermore, other than the personal computer platform, the application may be made available to the mobile platform rendering the application more ubiquitous. From the perspective of the user interface, the information can be displayed perfectly to increase the effectiveness not only through the 3D visualization but other multimedia elements to convey the content.

From the research goals, we expect two main contributions as follow:

- Promote the imaginative learning
- Obtain another experience of learning with the advance of AR technology.

Besides that it will benefit to all researchers, practitioners in AR, interaction design for children and Human Computer Interaction (HCI) in general.

Moreover, not only the content creation needs to be focused but the performance of the application through different

platform capabilities and performances. To reach the maximum effectiveness, the performances and error free also plays the important roles. For the developers of multimedia application and augmented reality system, the understanding of the tools will make the development process rapid but at the same time ensuring high quality of the final product.

Generally the most challenging part is the requirement analysis design for children; detailed consideration of each element must be taken into consideration especially for conceptual design in augmented reality applications. Another challenge has been in choosing the suitable tool for development of the prototype for the application so that it was completed within the time frame. As a conclusion, the project was successfully designed based on the design stages of our methodology.

REFERENCES

- [1] E. Barba and B. MacIntyre, "A scale model of mixed reality," *Proc. 8th ACM Conf. Creat. Cogn. - C&C '11*, p. 117, 2011
- [2] M. Inaba, A. Banno, T. Oishi, and K. Ikeuchi, "Achieving robust alignment for outdoor mixed reality using 3D range data," *Proc. 18th ACM Symp. Virtual Real. Softw. Technol. - VRST '12*, p. 61, 2012.
- [3] K. T. W. Tan, E. M. Lewis, N. J. Avis, U. Kingdom, and U. Kingdom, "Using augmented reality to promote an understanding of materials science to school children," 2008.
- [4] S. Siltanen and M. Aikala, "Augmented reality enriches hybrid media," *Proceeding 16th Int. Acad. MindTrek Conf. - MindTrek '12*, p. 113, 2012.
- [5] S. Livatino, F. Privitera, S. Superiore, and S. Visualization, "3D Visualization Technologies for Teleguided Robots," pp. 240–243, 2006.
- [6] A. Oulasvirta, S. Estlander, and A. Nurminen, "Embodied interaction with a 3D versus 2D mobile map," *Pers. Ubiquitous Comput.*, vol. 13, no. 4, pp. 303–320, Jul. 2008.
- [7] C. Healey, "Perceptually-motivated graphics, visualization and 3D displays" SIGGRAPH '10 ACM SIGGRAPH 2010 Courses (p. 37). New York: ACM. Igarashi, T. (July, 2010). Computer graphics for all. Communications of the ACM, Volume 53(7), 71-77.
- [8] R. Agrusa, V. G. Mazza, and R. Penso, "Advanced 3D visualization for manufacturing and facility controls," *2009 2nd Conf. Hum. Syst. Interact.*, pp. 456–462, May 2009.
- [9] G. H. Bendels, F. Kahlesz, and R. Klein, "Towards the next generation of 3D content creation," *Proc. Work. Conf. Adv. Vis. interfaces - AVI '04*, p. 283, 2004.
- [10] T. Olsson, E. Lagerstam, T. Kärkkäinen, and K. Väänänen-Vainio-Mattila, "Expected user experience of mobile augmented reality services: a user study in the context of shopping centres," *Pers. Ubiquitous Comput.*, vol. 17, no. 2, pp. 287–304, Dec. 2011.
- [11] R. T. Azuma, "A Survey of Augmented Reality," *Presence: Teleoperators and Virtual Environments* 6, no. 4 pp. 355–385, 1997.
- [12] H. B.-L. Duh and M. Billinghurst, "Trends in augmented reality tracking, interaction and display: A review of ten years of ISMAR," *2008 7th IEEE/ACM Int. Symp. Mix. Augment. Real.*, pp. 193–202, Sep. 2008.
- [13] Y. Rogers, H. Sharp and J. Preece, "Interaction Design: Beyond Human - Computer Interaction," Chichester, Wiley, 2011.
- [14] J. F. Franco, R. D. D. Lopes, A. Prof, L. Gualberto, P. Cep, E. Municipal, D. E. Fundamental, E. Silva, and B. Emef, "Three-dimensional digital environments and computer graphics influencing k-12 individuals digital literacy development and interdisciplinary lifelong learning," *Educators Program*. New York 2009.
- [15] E. Woods, M. Billinghurst, and D. Brown, "Augmenting the Science Centre and Museum Experience," no. Figure 3, pp. 230–236, 2003.
- [16] J. Beheshti, A. Large, C.-A. Julien, and M. Tam, "A comparison of a conventional taxonomy with a 3D visualization for use by children," *Proc. Am. Soc. Inf. Sci. Technol.*, vol. 47, no. 1, pp. 1–9, Nov. 2010.
- [17] F. Garzotto and M. Forfori, "Hyperstories and social interaction in 2D and 3D edutainment spaces for children," *Proc. seventeenth Conf. Hypertext hypermedia - HYPERTEXT '06*, p. 57, 2006.
- [18] Z. Zhou, A. D. Cheok, J. Pan, and Y. Li, "Magic Story Cube: an Interactive Tangible Interface for Storytelling," in *Proceedings of the 2004 ACM SIGCHI International Conference on Advances in computer entertainment technology*, 2004, pp. 364–365.

Human Activity Recognition for Surveillance Applications

Ahmed Taha, Hala H. Zayed

Computer Science Dept.
Faculty of Computers & Informatics, Benha University
{ahmed.taha, hala.zayed}@fci.bu.edu.eg

M. E. Khalifa and El-Sayed M. El-Horbaty

Basic Science Dept., Computer Science Dept.
Faculty of Computer & Information Sciences, Ain Shams University
{esskhalifa, shorbaty}@cis.asu.edu.eg

Abstract—The analysis of human activities is one of the most interesting and important open issues for the automated video surveillance community. In order to understand the behaviors of humans, a higher level of understanding is required, which is generally referred to as activity recognition. While traditional approaches rely on 2D data like images or videos, the development of low-cost depth sensors created new opportunities to advance the field. In this paper, a system to recognize human activities using 3D skeleton joints recovered from 3D depth data of RGB-D cameras is proposed. A low dimensional descriptor is constructed for activity recognition based on skeleton joints. The proposed system focuses on recognizing human activities not human actions. Human activities take place over different time scales and consist of a sequence of sub-activities (referred to as actions). The proposed system recognizes learned activities via trained Hidden Markov Models (HMMs). Experimental results on two human activity recognition benchmarks show that the proposed recognition system outperforms various state-of-the-art skeleton-based human activity recognition techniques.

Keywords— Activity Recognition; Depth Images; HMM; Behavior Analysis; Video Surveillance

I. INTRODUCTION

Video surveillance has attracted a lot of attention of the computer vision community in recent years. The increasing demand for safety and security has resulted in more research in intelligent surveillance. It has a wide range of applications, such as observing people in large waiting rooms, shopping centers, hospitals, eldercare, home-nursing, campuses or monitoring vehicles inside/outside cities, on highways, bridges, in tunnels etc. [1]. Currently, there is an increasing desire and need in video surveillance applications to be able to analyze human behaviors. Behavior analysis involves the analysis and the recognition of motion patterns to produce a high-level description of actions and interactions among objects [2]. Despite significant research efforts over the past few decades, action recognition remains a highly challenging problem. The difficulties of action recognition come from several aspects [3, 4]. Firstly, human motions are represented in a very high dimensional space. Moreover, interactions among different subjects complicate searching in this space. Secondly, performing similar or identical activities by different subjects exhibit substantial variations. Thirdly, visual data from traditional video cameras can only capture projective information of the real world, and are sensitive to lighting conditions.

The problem of behavior analysis is addressed under different terms. In the literature, action recognition and activity

recognition are the most common used terms [2, 5]. The term action is often confused with the term activity. Action usually refers to a sequence of primitive movements carried out by a single object, that is, an atomic movement that can be described at the limb level [5], such as a walking step. However, activity contains a number of sequential actions. i.e., dancing activity consists of successive repetitions of several actions, e.g. walking, jumping, waving hand, etc. Actions can be placed on a lower level than activities. Approaches for recognizing activities are often hierarchical in nature. They use previously recognized actions as their input. Different approaches are used to recognize low-level actions [6]. Some approaches use every single frame (2D templates, 3D object models), while others look at the entire video (spatio-temporal filtering, sub-volume matching). These techniques extract features and match them to a template in order to recognize an action. Other techniques, such as hidden Markov models (HMMs), estimate a model on the temporal dynamics of an action. The model parameters are learned from training data.

One of the most common methods for representing human action is the use of human's skeletal information. In the past, extracting accurate skeletal information from video streams was very difficult and unreliable, especially for arbitrary human poses. In contrast, motion capture systems could provide very accurate skeletal information of human actions based on active or passive markers positioned on the body [7]. However, the data acquisition was limited to controlled indoor

environments. Hence, skeletal-based recognition methods became less popular over the years as compared to the image feature-based recognition methods [7]. The latter methods extract spatiotemporal interest points from video images and the recognition is based on learned statistics on large datasets. Lately, new technologies help to enhance the monitoring process creating systems that are more powerful in detecting dangerous situations. With the release of several low-cost 3D capturing systems, such as the Microsoft Kinect, real time 3D data acquisition and skeleton extraction have become much easier and more practical for action recognition, thus restoring interest in the skeleton-based action recognition.

In this paper, a system for human activity recognition is proposed. Actually, we extend our previous work presented in [8] by focusing on recognizing complex activities as a sequence of basic actions. The proposed method presents a human activity descriptor based on the human's skeletal information extracted from Microsoft Kinect. This representation of the human activity is invariant to the scale of the subjects/objects and the orientation to the camera, while it maintains the correlation among different body parts. Hidden Markov Models (HMMs) are employed to recognize human activities. For each activity class, a HMM is learned. In the classification step, an unknown activity descriptor is aligned with the HMM in each class. An unknown sequence will be classified into the class, which has the highest alignment score.

The remainder of this paper is organized as follows: Section II gives a brief review of some related work in human activity recognition. In Section III, an overview of RGB-D sensor and depth images is provided. Section IV then presents the proposed system. The performance analysis of the proposed system is empirically evaluated in Section V. Finally, we conclude in Section VI.

II. RELATED WORK

Over the past decade, a great deal of work has been done on the recognition of human activities. However, the problem is still open and provides a big challenge to the researchers and more rigorous research is needed to come around it. An overview of the various action recognition methods and available well-known action datasets are provided in [9]. Most previous research in action recognition was based on color or greyscale intensity images. These images are obtained from traditional RGB cameras, where the value of each pixel represents the intensity of incoming light. It contains rich texture and color information, which is very useful for image processing, however it is very sensitive to illumination changes.

Recently, there have been vision technologies that can capture distance information from the real world, which cannot be obtained directly from an intensity image. These images are obtained from depth cameras, where the value of each pixel represents the calibrated distance between camera and scene. An advantage of using these sensors is that they give depth at every pixel so the shape of the object can be measured. When using depth images, computer vision tasks like background subtraction and contour detection become easier. Actually,

there are many attractive progresses and improves have been done with the use of depth information.

Based on the above, there are two main approaches for human behavior recognition: RGB video-based approach [9] and depth map-based approach [3, 4]. In this section, we focus only on reviewing the state-of-the-art techniques that investigate the applicability and benefit of depth sensors for action recognition especially skeleton-based approaches. The use of the different data provided by the RGB-D devices for human action recognition goes from employing only the depth data, or only the skeleton data extracted from the depth, to the fusion of both the depth and the skeleton data. Existing skeleton-based human action recognition approaches can be broadly grouped into two main categories [10]: joint-based approaches and body part-based approaches. Joint-based approaches consider human skeleton as a set of points, whereas body part-based approaches consider human skeleton as a connected set of rigid segments. Approaches that use joint angles can be classified as body part-based approaches since joint angles measure the geometry between directly connected pairs of body parts.

Jalal et al. [11] present a depth-based life logging human activity recognition system to recognize the daily activities of elderly people and turn these environments into an intelligent living space. Initially, a depth imaging sensor is used to capture depth silhouettes. Based on these silhouettes, human skeletons with joint information are produced which are further used for activity recognition and generating their life logs. The life-logging system is divided into two processes. Firstly, the training system includes data collection using a depth camera, feature extraction and training for each activity via Hidden Markov Models. Secondly, after training, the recognition engine starts to recognize the learned activities and produces life logs.

Gasparri et al. [12] propose a method for automatic fall detection using the Kinect depth sensor in top-view configuration. Their approach allows detecting a fall event without relying on wearable sensors, and by exploiting privacy-preserving depth data only. Starting from suitably preprocessed depth information, the system is able to recognize and separate the still objects from the human subjects within the scene using an ad-hoc discrimination algorithm. Several human subjects may be monitored through a solution that allows simultaneous tracking. Once a person is detected, he is followed by a tracking algorithm between different frames. The use of a reference depth frame, containing the set-up of the scene, allows one to extract a human subject, even when he/she is interacting with other objects, such as chairs or desks.

Althloothia et al. [13] present two sets of features for human activity recognition using a sequence of RGB-D images: shape representation and kinematic structure. The shape features are extracted using the depth information in the frequency domain via spherical harmonics representation. The other features include the motion of the 3D joint positions (i.e. the ends of the distal limb segments) in the human body. Both sets of features are fused using the Multiple Kernel Learning

(MKL) technique at the kernel level for human activity recognition.

Wang et al. [14] present an Actionlet Ensemble Model for human action recognition with depth cameras. An actionlet is a particular conjunction of the features for a subset of the joints, indicating a structure of the features. As there are an enormous number of possible actionlets, the authors propose a data mining solution to discover discriminative actionlets. Then an action is represented as an Actionlet Ensemble, which is a linear combination of the actionlets, and their discriminative weights are learnt via a multiple kernel learning method.

Oflin et al. [7] propose a skeletal motion feature representation of human actions, called Sequence of the Most Informative Joints (SMIJ). Specifically, in the SMIJ representation, a given action sequence is divided into a number of temporal segments. Within each segment, the joints that are deemed to be the most informative are selected. The sequence of such most informative joints is then used to represent an action. One of the limitations of the SMIJ representation that remains to be addressed is its insensitivity to discriminate different planar motions around the same joint. The joint angles are computed between two connected body segments in 3D spherical coordinates, thus capturing only a coarse representation of the body configuration.

III. RGB-D SENSOR

The Kinect sensor is a motion-sensing device that offers a simple and convenient way to capture and record features of human body motion [15]. The Kinect sensor produces a new type of data, RGB-D data, which is an improvement on RGB images for human behavior recognition research. Its name is a combination of kinetic and connects [16]. It was initially used as an input device by Microsoft for the Xbox game console. All user movements are captured and reflected on-screen. It enables the user to interact and control software on the Xbox 360 with gestures recognition and voice recognition. The Kinect's output is a multi-modal signal, which gives RGB videos, depth sequences and skeleton information simultaneously. Recently, the computer vision community discovered that the depth sensing technology of Kinect could be extended far beyond gaming and at a much lower cost than traditional 3D cameras (such as stereo cameras and Time-Of-Flight cameras) [17].

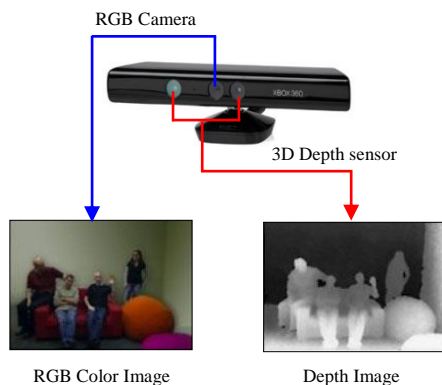


Fig. 1 RGB-D data captured by Kinect

Figure 1 shows the Kinect sensor and the RGB-D data captured including both RGB color image and depth image. A depth image (or depth map) is an image that contains information relating to the distance of the surfaces of scene objects from a viewpoint [18]. Pixels in a depth image indicate calibrated depth in the scene, rather than a measure of intensity or color. The device is actually composed of multiple sensors. In the middle, it has a RGB camera allowing a resolution up to 1280×960 at 12 images per second [16]. The usual used resolution is 640×480 pixels at 30 images per second maximum for colored video stream as the depth camera has a maximum resolution of 640×480 at 30 frames per second. A little away on the left of the device, It has the IR light (projector). It projects multiple dots, which allow the final camera on the right side, the CMOS depth camera, to compute a 3D environment. The device is mounted with a motorized tilt to adjust the vertical angle.

One of the major components of the Kinect sensor is its ability to infer human motion by extracting human silhouettes in skeletal structures. It extracts the skeletal joints of a human body as 3D points using the Microsoft SDK. It provides a skeleton model with 20 joints as shown in Figure 2. The complementary nature of the depth and visual RGB information provided by Kinect initiates new solutions for classical problems in computer vision. The availability of depth information allows researchers to implement simpler identification procedures to detect human subjects. The advantages of this technology, with respect to classical video-based ones, are [12]:

- Being less sensitive to variations in light intensity and texture changes;
- Providing 3D information by a single camera, while a stereoscopic system is necessary in the RGB domain to achieve the same goal;
- Maintaining privacy, it is not possible to recognize the facial details of the people captured by the depth camera. This feature helps to keep identity confidential.

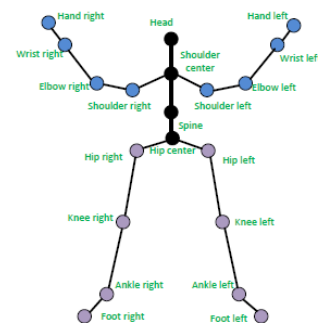


Fig. 2 [15] Skeleton joints detected by Microsoft SDK

IV. PROPOSED SYSTEM

The proposed method focuses on obtaining a descriptive labeling of the complex human activities that take place over

different time scales and consist of a sequence of sub-activities (actions). In fact, human activity recognition is a challenging task since it needs to face with numerous varieties. First, the variation in the length of an action where different individuals perform actions at diverse rate. Second is differences in the characteristics of the human body such as body shape, height, weight fitting, etc. Third is the ambiguity caused by the similarity of some activities, which represents a great challenge for any recognition system. Moreover, environment settings and video quality should be considered. For example, dynamic backgrounds and cluttered environments are always difficult to handle in any video processing application. Other factors such as lighting condition, camera viewpoint, and camera motion should also be addressed properly.

activity recognition. These activities take place over a long period and consist of a sequence of sub-activities. The proposed system employs the human action representation presented in [8] to recognize complex activities. This representation is characterized by its low dimensionality and its invariance to the scale of the subjects/objects and the orientation to the camera, while it maintains the correlation among different body parts. It is based on the human's skeletal information extracted from depth images. The basic idea of the proposed system depends on the fact that each activity consists of a sequence of sub-activities (actions) that change over the course of performing the activity. For example, a suspicious activity like leaving a bag in a public place may include the suspect walk, bend and run in a sequence. Therefore, the proposed system recognizes these actions independently. Then, an activity descriptor is constructed from these actions as an ordered sequence. Initially, the descriptor is empty. Then, every detected action is added in order to the sequence. Later, trained Hidden Markov Models (HMMs) are used for recognizing unknown activities.

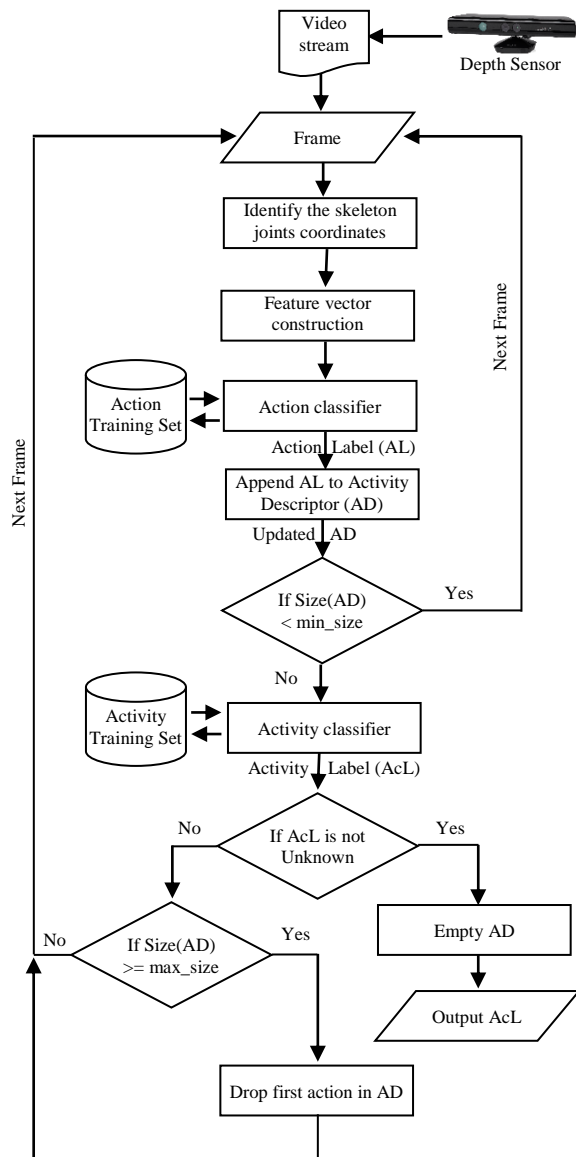


Fig. 3 The block diagram of the proposed system

In fact, our previous work in [8] focuses on recognizing actions that span short time periods. However, in this paper, the proposed system extends that work by performing a high-level

Figure 3 shows the block diagram of the proposed system. First, the system starts with identifying the skeleton joints coordinates for each detected object in the video sequence. Actually, the Kinect camera tracks 20 body joints for each object in the scene. The position of the skeleton joints are provided as Cartesian coordinates (X, Y, Z) with respect to a coordinate system centered at the Kinect. The positive Y axis points up, the positive Z axis points where the Kinect is pointing, and the positive X axis is to the left as shown in Figure 4.

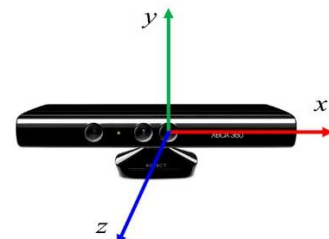
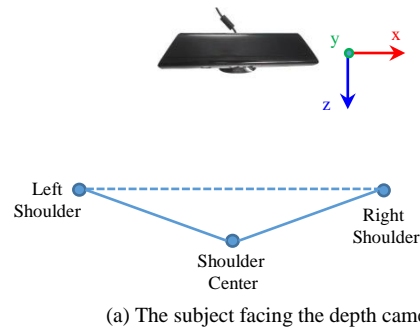


Fig. 4 Kinect Cartesian coordinate system



(a) The subject facing the depth camera

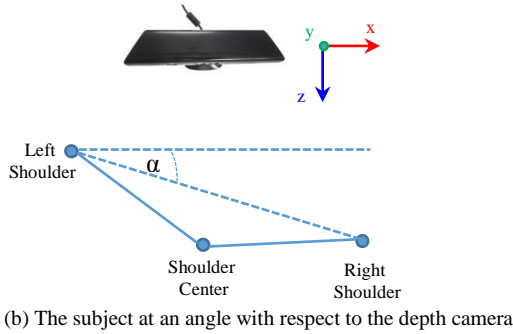


Fig. 5 Rotation of the skeleton with respect to the Kinect

Second, the proposed system constructs the feature vector for each detected skeleton in the scene. Ideally, a subject should be straight in front of Kinect camera (Figure 5.a) but this is not always the case. The subject can be at any angle from Kinect (Figure 5.b) and at any distance. To overcome this issue, the proposed system rotates all the skeleton points around Y-axis in a counterclockwise direction with an angle α in order to make the subject straight in front of depth camera. Hence, rotation invariance is achieved. This angle is defined as the angle between the line connecting both shoulders and the positive direction of X-axis of Kinect coordinates system (Figure 5.b). Initially, the angle α is estimated using the coordinates of two joints: shoulder left (x_L, y_L, z_L) and shoulder right (x_R, y_R, z_R) as in (1):

$$\alpha = \tan^{-1} \left(\frac{z_R - z_L}{x_R - x_L} \right) \quad (1)$$

Then a counterclockwise rotation about Y-axis is applied to all skeleton joints with an angle α . For each skeleton joint i with coordinates (x_i, y_i, z_i) , the rotated coordinates (x'_i, y'_i, z'_i) are calculated using (2):

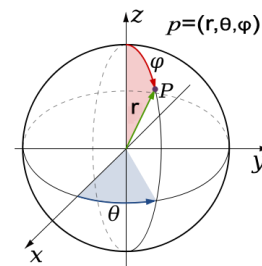
$$\begin{bmatrix} x'_i \\ y'_i \\ z'_i \\ 1 \end{bmatrix} = \begin{bmatrix} \cos \alpha & 0 & \sin \alpha & 0 \\ 0 & 1 & 0 & 0 \\ -\sin \alpha & 0 & \cos \alpha & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ z_i \\ 1 \end{bmatrix} \quad (2)$$

Moreover, varying the object distance from Kinect makes the action recognition more sophisticated. Therefore, it is necessary to shift the origin of the coordinates from Kinect to a point in the object body to remove dependence on camera position. This means joints coordinates should be translated to another coordinate system where its origin is a point in the human body rather than the Kinect camera. By this way, the distance factor between the object and Kinect is neutralized. This permits the coordinates to be expressed invariantly to translation and rotation of the body with respect to the camera reference system. In our proposed system, we use the shoulder center joint as the origin of the new system (see Figure 2). Assume that shoulder center joint coordinates are (x, y, z) . Hence for each skeleton joint i with coordinates (x_i, y_i, z_i) , the translated coordinates (x'_i, y'_i, z'_i) are calculated with (3):

$$(x'_i, y'_i, z'_i) = (x_i - x, y_i - y, z_i - z) \quad (3)$$

Moreover, the individual variations of people in terms of posture, height and dimensions have a huge impact on the performance of the action recognition system. This is because X, Y and Z coordinates of joints of every object doing the same action might be different. Therefore, it is necessary to normalize the data to increase accuracy of action recognition. To simplify the normalization process, the joints coordinates are converted from Cartesian coordinate system to spherical coordinate system. The spherical coordinate system is a three dimensional space system with three components: the distance of the point from the origin (radial distance r), the polar angle (φ), and the azimuth angle (θ) as shown in Figure 6. When normalizing a point in Cartesian coordinates, all the components X, Y and Z are changed. However when normalizing a point in the spherical coordinates, only radial distance r will equal to one while both polar angle (φ) and azimuth angle (θ) will remain constant.

Feature vectors provide a set of characteristics that represent the action to be recognized. However, it may include irrelevant or redundant information which could complicate the classification. Reducing the feature vector size has an important impact on the processing time since the recognition is performed faster. Concerning the skeletal data obtained with depth sensor devices, it can be seen that some joints are more important than others if action recognition is targeted. Several joints in the torso (the skeleton part identified by a dashed line in Figure 7) do not show an independent motion along with the whole body. Hence, in our proposed system, seven joints coordinates of the human skeleton are discarded from the feature vector. These joints are shown as solid circles in Figure 7: shoulder right, shoulder center, shoulder left, spine, hip center, hip right, and hip left (from left-to-right and from top-to-bottom respectively). This dimensionality reduction of the feature vector improves the classification performance. Since the joints coordinates are normalized, radial distance r can be ignored in our feature vector. Thus, the feature vector will consist of 13 pairs of (φ, θ) for each detected object in the scene. This means it has only 26 components which is a reduced feature vector than what is reported in the state-of-the-art methods [19-21]. A low-dimensional representation means less computational effort.



$$r = \sqrt{x^2 + y^2 + z^2}, \quad \theta = \cos^{-1} \left(\frac{z}{r} \right), \quad \varphi = \tan^{-1} \left(\frac{y}{x} \right)$$

Fig. 6 Spherical coordinates (r, θ, ϕ): radial distance r , azimuthal angle θ , and polar angle ϕ

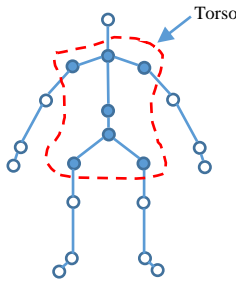


Fig. 7 Torso skeleton joints discarded from the feature vector

After a feature vector is constructed, a classification step is needed to recognize different actions. The feature vector of the unknown action is used as input to the classifier whose objective is to accurately identify which action class is best matched against the input. In our proposed system, a Multi-class Support Vector Machine (MSVM) [22-24] is employed to perform action classification. The MSVM used is based on One-Against-All (OAA) classification approach [23] where there is one binary SVM for each class to separate members of that class from members of other classes. A data point would be classified under a certain class if and only if that class's SVM accepted it and all other classes' SVMs rejected it. A training step is needed to summarize the similarity within (and dissimilarity in-between) the training samples of different action classes. With action models learned, a new action instance can be recognized as one of the learned classes.

Once an action is recognized, it is a candidate to be a part of a more complex activity. This is because a human activity is actually a series of human actions. In order to recognize this activity, the proposed system constructs and maintains an activity descriptor. It is simply an ordered list of the detected actions and it satisfies two criteria. First, adjacent actions in the activity descriptor are not allowed to be the same. However, the activity descriptor may contain the same action more than one time but not adjacent. Second, the activity descriptor is variable length with a special notion of order since not all activities consist of the same number of actions. However, a minimum and a maximum size of the descriptor is initially predetermined from the training set. Initially, the activity descriptor is an empty set and it is updated each time either an action or an activity is recognized.

Considering the nature of the proposed activity descriptor, the problem of recognizing activities can be formulated as a sequence classification problem. Given L as a set of class labels, the task of sequence classification is to learn a sequence classifier C , which is a function mapping of a sequence s to a class label $l \in L$, written as, $C : s \rightarrow l; l \in L$. In the proposed system, HMMs are employed for performing action recognition, due to their suitability for modeling pattern

recognition problems that exhibit an inherent temporality. HMMs are one of the most popular generative models used for classification. It is a doubly stochastic process [25]. The underlying stochastic process is not observable but can be observed through another set of stochastic processes that produce the sequence of observed symbols [25]. The underlying hidden stochastic process is a first-order Markov process; that is, each hidden state depends only on the previous hidden state. Moreover, in the observed stochastic process, each observed measurement (symbol) depends only on the current hidden state. The use of HMMs includes two stages: learning and recognition. In the learning stage, the data are used to optimize the parameters of the HMM of each activity (class). That is, it involves developing a model for all of the activities that we want to recognize. In the recognition stage, the HMM of each class computes the probability of generating a test sequence, and the model which has the maximum probability is chosen.

Back to Figure 3, when an action is recognized, the action is appended to the activity descriptor provided it does not match the last action in the descriptor. If the descriptor size is less than the minimum size, the proposed system will proceed to the next frame to detect more actions to be added to the descriptor. Otherwise, when the descriptor reaches the minimum size, it is a candidate to be an activity. At this point, the activity descriptor is checked against all the trained HMMs to calculate the likelihood and the one having highest probability is chosen. Thus, to test an activity descriptor sequence AD , the HMMs act as (4):

$$AcL = \arg \max_{i=1,2,\dots,N} \{P(AD|H_i)\} \quad (4)$$

where the activity label (AcL) is based on the probability of the activity descriptor (AD) on corresponding trained activity HMM H_i . When an activity is recognized, the proposed system resets the descriptor. It becomes empty again and ready for receiving more actions of the next activity. However, if the activity is not recognized, so the actions in the descriptor are not sufficient to recognize the activity. In this case, the descriptor size is checked against reaching to the maximum size. If so, the first action in the descriptor is dropped leaving the empty space for adding one more action. Otherwise, the proposed system proceeds to the next frame to recognize next actions.

V. EXPERIMENTAL RESULTS

In this section, experimental results of the proposed system are presented. Establishing standard test beds is a fundamental requirement to compare systems performance. There have been many human action benchmarks proposed in the literature (such as Weizmann, KTH and UCF datasets) [9]. Unfortunately, most of the existing benchmarks provide only color-based information but lack the corresponding depth data. However, with the advent of the Microsoft Kinect sensor, new 3D depth datasets have emerged for human motion tracking, pose estimation and action recognition, such as MSR-

Action3D dataset [26], MSR Daily Activity3D dataset [14], and Florence 3D Action dataset [27]. These datasets provide a rich depth representation of the scene at each time instant, allowing for both spatial and temporal analysis of human motion. To evaluate the performance of the proposed system, experiments were carried out on both MSR Daily Activity3D dataset and Florence 3D Action dataset while MSR-Action3D dataset is excluded. This is due to MSR-Action3D dataset contains just actions not activities so it is usually used for evaluating action recognition techniques.

MSR Daily Activity3D dataset [14] is a benchmark dataset used widely to evaluate the performance of RGBD-based activity recognition methods [7, 12-14]. It is a daily activity dataset captured by a Kinect device at Microsoft research. There are background objects and persons appearing at different distances to the camera. Also, this dataset is rather challenging because most of the activities involves human-object interactions. The dataset includes 320 samples from sixteen different action classes, and for each sample, depth sequence, RGB video and skeleton information are provided. The activity types include: drink, eat, read book, call cellphone, write on a paper, use laptop, use vacuum cleaner, cheer up, sit still, toss paper, play game, lay down on sofa,

walk, play guitar, stand up and sit down. Each subject performs each activity twice, once in standing position, and again in sitting on sofa position. Figure 8 gives some example frames of MSR Daily Activity3D dataset. The first row shows RGB frames while the second row shows their corresponding depth images extracted from Kinect sensor. The full dataset can be downloaded from (<http://research.microsoft.com/en-us/people/zliu/ActionRecoRsrc/default.htm>).

The second dataset used in the experiments is Florence 3D Action dataset. It is collected at the University of Florence during 2012 and it has been captured using a Kinect camera. It includes nine activities: wave, drink from a bottle, answer phone, clap, tight lace, sit down, stand up, read watch and bow. During acquisition, 10 subjects were asked to perform the above actions for two or three times. This resulted in a total of 215 activity samples. The main challenges of this dataset are the similarity between actions, the human object interaction, and the different ways of performing the same action. Figure 9 shows some example frames of Florence 3D Action dataset. Each column shows an activity performed by three different subjects. The full dataset can be downloaded from (<http://www.micc.unifi.it/vim/datasets/3dactions/>).

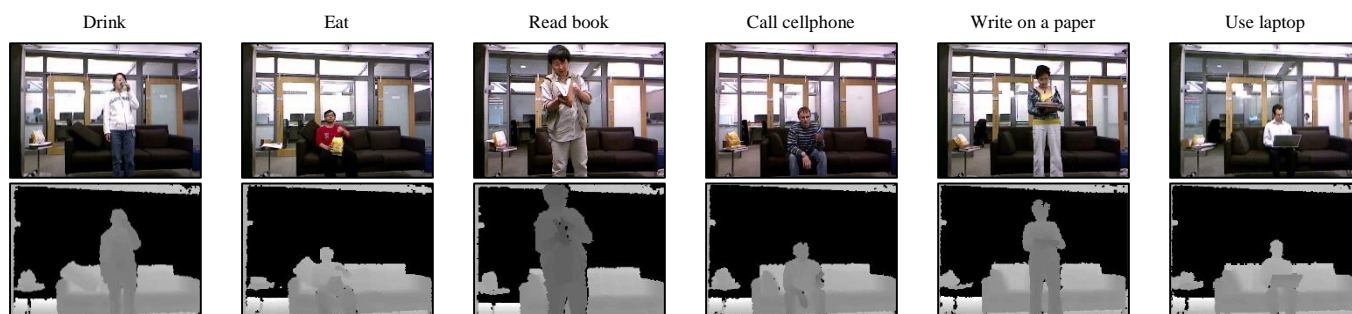


Fig. 8. Some example frames of MSR Daily Activity3D dataset, First row: RGB frames, Second row: depth images

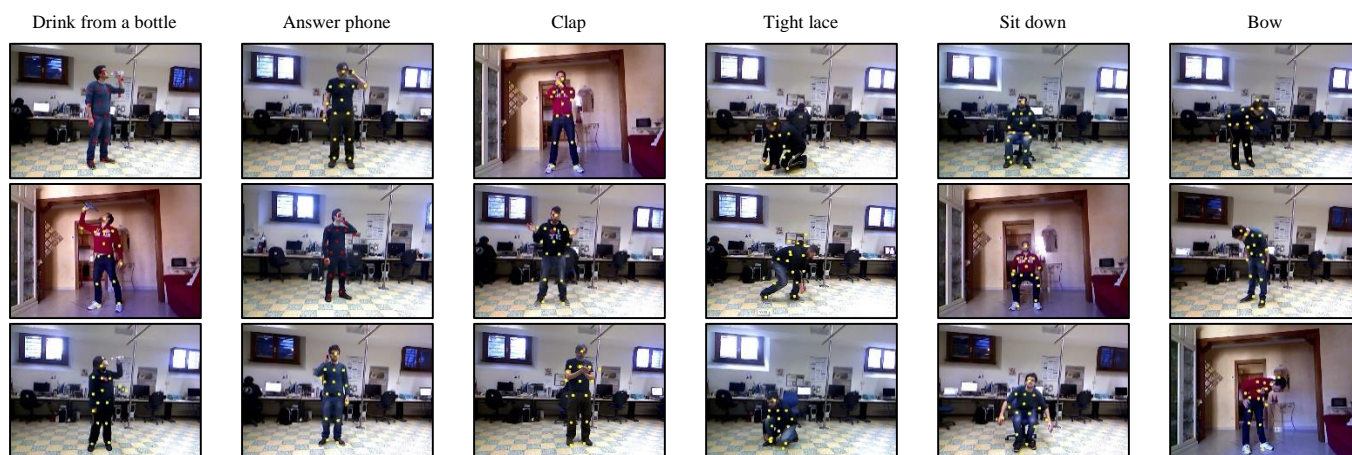


Fig. 9. Some example frames of Florence 3D dataset, each activity is performed by different subjects

- drink
- eat
- Read Book
- Call Cellphone
- write
- Use Laptop
- Vacuum Cleaner
- cheerUp
- sitStill
- tossPaper
- Play Game
- layDown
- walk
- Play Guitar
- standUp
- sitDown

drink	96	2					2												
eat	6	91																3	
readBook			85		9													6	
callCellphone	11	4		79			6												
write					88			5		7									
useLaptop			3		2	94								1					
vaccumCleaner							100												
cheerUp								100											
sitStill									100										
tossPaper				3		4				92				1					
playGame			3			1			7		89								
layDown												99							1
walk													100						
playGuitar														100					
standUp																		100	
sitDown												2							98

Fig. 10. The confusion matrix of the proposed system on Daily-Activity3D dataset

It should be mentioned that all experiments were implemented on a 2.5GHz Intel Core i7 PC with 4GB memory, running under Windows 8 Enterprise. The proposed system is coded using MATLAB 8.1.0.604 (R2013a). During the experiments, we used a cross-subject training/testing setup in which we take out each subject (i.e., leave-one-subject-out scheme) from the training set and repeat an experiment for each of them. This is the same settings used in evaluating the state-of-the-art methods [11, 13, 14, 27]. Figure 10 and Figure 11 show the confusion matrices of the proposed system using MSR Daily Activity3D dataset and Florence 3D dataset respectively.

Each row represents the instances in an actual class and each column denotes the recognition results. For example in the second row of Figure 10, 91% of the “eat” samples are classified correctly while 6% of the samples are misclassified as “drink” activity and 2% are misclassified as “play guitar” activity. As, it can be seen from the figure, the results prove the efficiency of the proposed method in recognizing different activities.

	wave	drink	answer	clap	tight	sitdown	standup	read watch	bow
wave	99			1					
drink		98	2						
answer	2	4	94						
clap				100					
tight					100				
sitdown				2		97			1
standup				1	2		94		3
read watch	1			3				95	
bow		1	2			4	2		89

Fig. 11. The confusion matrix of the proposed system on Florence 3D Action dataset

Moreover, we compare the performance of the proposed system with several recent methods [11, 13, 14, 27] and summarize the results in Table I. It is clear that the proposed system outperforms the other approaches on both MSR Daily Activity3D and Florence 3D benchmarks. It can be also noted that the recognition accuracies achieved for Florence 3D benchmark are better than those for MSR Daily Activity3D benchmark. This is because the Florence 3D Action dataset has fewer classes than MSR Daily Activity3D and action samples are shorter on average. The Florence3D dataset is

probably less difficult than MSR Daily Activity3D because only a few activities are performed through external object interactions.

Furthermore, we can see that the results achieved by Seidenari et al. [27] are the lowest recognition accuracies comparing to the other methods. The main reason for their low accuracy is that their work aims to show the powerful of information that can be extracted from the 3D skeleton only, without requiring the additional processing of the entire depth maps of a sequence. In addition, Jalal et al. [11] suffers from the high dimensionality of the motion parameter vectors used to represent joint points features. This drawback incurs more complexity to their work. Also, the Actionlet method proposed by Wang et al. [14] uses high ordering features and complicated learning procedures that limit its use in real time applications. The multi-fused features method proposed by Althloothia et al. [13] uses large-dimensionality features, which needs high computational times that make it impractical for long-term human action recognition and real-time applications. Meanwhile, our proposed system is quite simple for computation purposes and provides sufficient and compact feature information.

TABLE I. RECOGNITION ACCURACIES (%) OF THE PROPOSED SYSTEM COMPARED TO THE STATE-OF-THE-ART METHODS

Method	Datasets	
	MSR Daily Activity3D	Florence 3D
Wang et al. (2012) [14]	85.7%	NA
Seidenari et al. (2013) [27]	70%	82%
Jalal et al. (2014) [11]	79.1%	NA
Althloothia et al. (2014) [13]	93.1%	NA
The proposed system	94.4%	96.2%

VI. CONCLUSION AND FUTURE WORK

Recently, with the availability of inexpensive RGB-D sensors, the problem of human activities recognition has become relatively easier and more robust. However, most of

these works only address detecting actions that stretches over short time periods not activities. In this paper, a system for human activity recognition is proposed. We have considered the task of obtaining a descriptive labeling of the activities being performed through labeling human sub-activities. The activities we consider happen over a long period, and comprise several sub-activities performed in a sequence. The proposed activity descriptor makes the activity recognition problem viewed as a sequence classification problem. The proposed system employs Hidden Markov Models (HMMs) to recognize human activities. Experiments carried out on two benchmark datasets support the applicability of the proposed solution. When compared to other skeletal-based solution our approach shows competitive performance.

As a future work, we would like to apply our proposed system to recognize human activities during a large amount of time. We may also extend this work for healthcare monitoring system, where the activities of patients are important for research.

REFERENCES

- [1] Kavita V. Bhaltilak, Harleen Kaur, Cherry Khosla, "Human Motion Analysis with the Help of Video Surveillance: A Review," In the International Journal of Computer Science Engineering and Technology (IJCSSET), Volume 4, Issue 9, pp. 245-249, September 2014.
- [2] Chen Change Loy, "Activity Understanding and Unusual Event Detection in Surveillance Videos," PhD dissertation, Queen Mary University of London, 2010.
- [3] Mao Ye, Qing Zhang, Liang Wang, Jiejie Zhu, Ruigang Yang, Juergen Gall, "A Survey on Human Motion Analysis from Depth Data," Lecture Notes in Computer Science, Springer Berlin Heidelberg, Volume 8200, pp 149-187, 2013.
- [4] Lulu Chen, Hong Wei, James Ferryman, "A survey of human motion analysis using depth imagery," In Pattern Recognition Letters, Elsevier Science Inc., Volume 34, Issue 15, pp. 1995-2006, November 2013.
- [5] Ronald Poppe, "A survey on vision-based human action recognition," In the International Journal of Image and Vision Computing, Volume 28, Number 6, pp.976-990, June 2010
- [6] Maaïke Johanna, "Recognizing activities with the Kinect," Master thesis, Radboud University Nijmegen, Nijmegen, Netherlands, July 2013.
- [7] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy, "Sequence of the Most Informative Joints (SMIJ): A New Representation for Human Skeletal Action Recognition," In proceedings of the IEEE Computer Vision and Pattern Recognition Workshops (CVPRW), Providence, Rhode Island, USA, PP. 8-13, June 2012.
- [8] Ahmed Taha, Hala H. Zayed, M. E. Khalifa and El-Sayed M. El-Horbaty, " Human Action Recognition based on MSVM and Depth Images," In The International Journal of Computer Science Issues (IJCSI), Volume 11, Issue 4, Number 2, pp. 42-51, July 2014.
- [9] Ahmed Taha, Hala H. Zayed, M. E. Khalifa and El-Sayed M. El-Horbaty, "Exploring Behavior Analysis in Video Surveillance Applications," In The International Journal of Computer Applications (IJCA), Foundation of Computer Science, New York, USA, Volume 93, Number 14, pp. 22-32. May 2014.
- [10] Raviteja Vemulapalli, Felipe Arrate and Rama Chellappa, "Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group," In Proceedings of the International IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, Ohio, USA, pp.588-595, June 2014.
- [11] Ahmad Jalal, Shaharyar Kamal and Daijin Kim, "A Depth Video Sensor-Based Life-Logging Human Activity Recognition System for Elderly Care in Smart Indoor Environments," In the International Journal of Sensors, Volume 14, Number 7, pp. 11735-11759, July 2014.
- [12] Samuele Gasparrini, Enea Cippitelli, Susanna Spinsante and Ennio Gambi, "A Depth-Based Fall Detection System Using a Kinect Sensor," In the International Journal of Sensors, Volume 14, Issue 2, pp. 2756-2775, February 2014.
- [13] Salah Althloothia, Mohammad H. Mahoor, Xiao Zhanga, Richard M. Voylesb, "Human Activity Recognition Using Multi-Features and Multiple Kernel Learning," In Pattern Recognition Journal, Volume 47, Issue 5, pp. 1800-1812, May 2014.
- [14] Jiang Wang, Zicheng Liu, Ying Wu, Junsong Yuan, "Mining Actionlet Ensemble for Action Recognition with Depth Cameras," In Proceedings of the International IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, Rhode Island, USA, pp. 1290-1297, June 2012.
- [15] Xiaoxiao Dai, "Vision-based 3D Human Motion Analysis for Fall Detection and Bed-exiting," Master thesis, Faculty of the Daniel Felix Ritchie School of Engineering and Computer Science, University of Denver, USA, August 2013.
- [16] Manjuatha M B, Pradeep kumar B.P., Santhosh.S.Y, "Survey on Skeleton Gesture Recognition Provided by Kinect," In the International Journal of Advanced Research in Electrical Electronics and Instrumentation Engineering (IAREEIE), Volume 3, Issue 4, April 2014.
- [17] Jungong Han, Ling Shao, Dong Xu, and Jamie Shotton, "Enhanced Computer Vision with Microsoft Kinect Sensor: A Review," In IEEE Transactions on Cybernetics, Volume 43, Number 5, pp. 1318 - 1334, October 2013.
- [18] Vennila Megavannan, Bhuvnesh Agarwal, and R. Venkatesh Babu, "Human Action Recognition using Depth Maps," In proceedings of the International Conference on Signal Processing and Communications (SPCOM), Bangalore, India, pp. 1-5, July 2012.
- [19] Alexandros Andre Charaouia, José Ramón Padilla-López, Pau Climent-Pérez, and Francisco Flórez-Revuelta, "Evolutionary Joint Selection to Improve Human Action Recognition with RGB-D Devices," In the International Journal of Expert Systems with Applications, Volume 41, Issue 3, pp. 786-794, February 2014.
- [20] Xiaodong Yang, Chenyang Zhang, and YingLi Tian, "Recognizing Actions Using Depth Motion Maps-Based Histograms of Oriented Gradients," In Proceedings of the 20th ACM International Conference on Multimedia (MM '12), New York, USA, pp. 1057-1060, November 2012.
- [21] Xiaodong Yang, and Yingli Tian, "EigenJoints-Based Action Recognition Using Naïve-Bayes-Nearest-Neighbor" In Proceeding of the International IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Providence, Rhode Island, USA, pp. 14-19, June 2012.
- [22] Xisheng He, Zhe Wang, Yingbin Zheng, and Xiangyang Xue, "A Simplified Multi-Class Support Vector Machine with Reduced Dual Optimization" In Pattern Recognition Letters Journal, Volume 33, Issue 1, pp. 71-82, January 2012.
- [23] Xiaowei Yang, Qiaozhen Yu, Lifang He, and Tengjiao Guo, "The One-Against-All Partition Based Binary Tree Support Vector Machine Algorithms for Multi-Class Classification," In the Neurocomputing Journal, Volume 113, pp. 1-7, August 2013.
- [24] Henry Joutsijoki, and Martti Juhola, "Kernel Selection in Multi-Class Support Vector Machines and its Consequence to the Number of Ties in Majority Voting Method," In Artificial Intelligence Review Journal, Volume 40, Issue 3, pp. 213-230, October 2013.
- [25] Shian-Ru Ke, Hoang Le Uyen Thuc, Yong-Jin Lee, Jenq-Neng Hwang, Jang-Hee Yoo, Kyoung-Ho Choi, "A Review on Video-Based Human Activity Recognition," In the International Journal of Computers, Volume 2, Issue 2, pp.88-131, June 2013.
- [26] Wanqing Li, Zhengyou Zhang, and Zicheng Liu, "Action Recognition Based on a Bag of 3D Points," In Proceedings of the IEEE International Computer Vision and Pattern Recognition Workshops (CVPRW), San Francisco, CA, pp. 9-14, June 2010.
- [27] Lorenzo Seidenari, Vincenzo Varano, Stefano Berretti, Alberto Del Bimbo, and Pietro Pala "Recognizing Actions from Depth Cameras as

A Comparative Study on The Existing Graphical User Interfaces for Occupational Therapy

Maryam Tayefeh Mahmoudi
Multimedia Research Group, IT Research Faculty
Research Institute for ICT
Tehran, Iran
mahmodi@itrc.ac.ir

Kambiz Badie
Knowledge Management & e-Organization Group
IT Research Faculty, Research Institute for ICT
Tehran, Iran
K_badie@itrc.ac.ir

Shahab Hossein Ahmadi Varnousfaderani
Department of Surveying Engineering
Faculty of Engineering, University of Tehran
Tehran, Iran
shahab_hossein@ut.ac.ir

Abstract—In this paper, we present a comparative study on the existing graphical user interfaces (GUIs) as well as software application peculiarities used for occupational therapy. In our approach we make use of the peculiarities belonging to GUIs themselves and the corresponding software applications. Here, a tabular form is used to show the functionalities of the existing GUIs in association with the corresponding peculiarities. Through the comparative study, an opportunity is gained to build a hybrid strategy for designing GUIs that can function in case of patients with mixed disorders. This is quite interesting since many patients are in fact suffering from a combination of disorders.

Keywords— *Graphical user interface (GUI); software application peculiarities; occupational therapy; physical disorder; mental disorder; tabular form; functionality.*

I. INTRODUCTION

In recent years, Graphical User Interfaces (GUIs) have been widely used in a wide range of Human Computer Interaction issues in the realms of education & research, decision-making, idea/art creation, medical diagnosis & treatment, movement/ posture control, etc. [1,2,3,4]. The point significant in all these cases is facilitating human computer interaction/ communication with the aim of guiding a computer and its peripheral systems in a direction compatible with the user's objectives. This means that, on the one side, computer should understand well what a user is intending, and on the other side, based on this understanding presents its decision to the user in terms of some messages or actions.

Among the issues calling for GUI, movement/ posture control in the disabled or patients with movement disorders is of particular significance. It is easily seen that, the message provided by GUI in this case, has the ability to give the patient an idea on how to select his/ her posture as well as control mental disorder. In this way, developing GUIs based on the

features of disabilities or disorders, may achieve a fruitful role in minimizing patient's stress/ anxiety, and besides that increasing the efficiency of the related treatment [4, 5]. Due to significance of such a notion, in this paper, we decided to make a comparative study on a variety of GUIs already developed for treating movement disorders, to figure out what peculiarities should exist in a GUI to make it most fittble for certain pathological situations. It is to be noted that, due to the difference in cognitive/ affective preferences of different patients with different types of disorders, the corresponding GUIs should be benefited by different peculiarities [6,7]. That is the main point we are trying to crystallize in this paper.

II. RELATED WORKS

There has been a growing interest in using technology platforms and software applications for training purposes in physical or mental occupational therapy. In these assisting software packages, the component which has a significant role in attracting audiences is Graphical User Interface (GUI).

Different perspectives may affect GUI design, such as: aesthetics, pedagogy, user-friendliness, cultural appropriateness, accessibility, effectiveness, engagement and innovation, etc. [8,9]. Beside the above-mentioned perspectives, in the case of Occupational Therapy, for those patients, their caregivers or therapists, who may not be able or familiar enough to apply those software packages, some other considerations may also be added. For instance, the GUI should be "simple", "user in control" and "consistent" to motivate the patient or his/her caregivers for continuous use.

Existing research activities in the domain of occupational therapy, reveals that mental-assisting software tools have a history longer than physical ones. The most feasible example of using such tools returns to supporting social and self-management skills in children with Autistic Spectrum Disorders (ASD), for which there is a burgeoning progress of designing assistant applications. Some of these applications focus on improving complete tasks, as well as transition within and between tasks [10]. While, the others present standardized measurement tools to estimate PDA's efficiency as cognitive assistants [11]. It is mentionable that, in addition to applications for patients, there also exist some applications for parents and instructors (ex. "HANDS") to train them how to assist the patients [12]. Also, in last years numerous mobile applications have been designed to help patients, suffering from autism including: "iOT session" [13], "Shelby's Quest" [14], "Brain Works" [15], "Autism Learning Games Camp Discovery" [16], and "Find me" [17], etc..

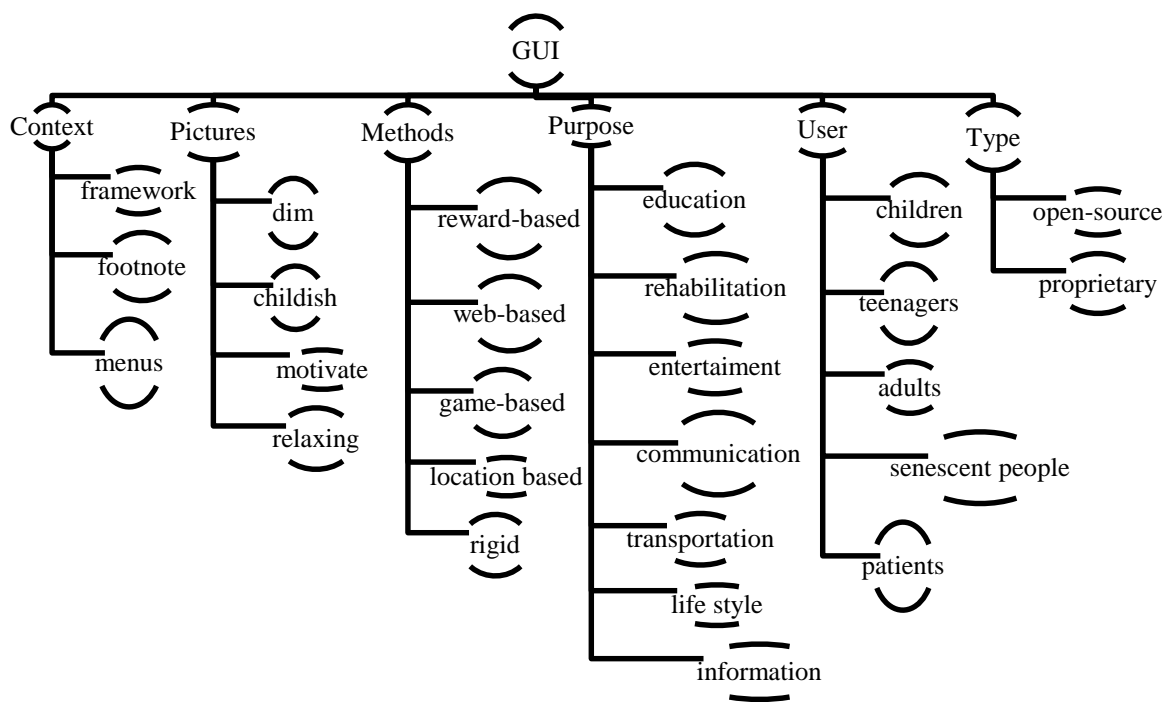
Another field of exploiting technology for mental therapy is assisting individuals with moderate-to-severe memory impairment, for which extensive fields of research have been done including: supporting memory function in individuals with milder memory impairment using mobile phones [18, 19], digital voice recorders [20] and PDAs and smart phones [21,

22]. In all these studies, the role of caregivers is crucial specifically in the cases where patients are unable to program the reminders, calendars and troubleshooting parts. Although, the first suggestion has been on structured training to make patients independent in using PDAs [23, 24, 25, 26], this suggestion is however not adequate for patient with severe memory impairment [26, 27]. Therefore a prosthetic memory protocol that teaches how to store and retrieve information in/ from a memory book, has been developed for patients with severe amnesia [28].

Beside the above-mentioned applications, there exist some software applications for other mental disturbances such as: depression, insomnia and anxiety. Some of the well-known related applications are: "Operation Reach Out" [29], "BellyBio" [30], "Optimism" [31], "iSleep Easy" [32], etc.

Moreover, physical disabilities have also been assisted by software applications in the present decade. For different types of pathological disorders, specific applications have been developed. For instance, "PT and OT helper"[33] and "Dexteria VPP"[34] are useful for fingers and hand exercises, while, "Occupational therapy" [35], "Therapy Boss" [36], "Physical Therapy Home exercises" [37] and "Physical Therapy for kids" [38], cover various parts of body exercise even for daily activities.

Taking the above points into account, the application and its graphical user interface should be designed using a variety of parameters (Fig.1), such as context, method, purpose, user and type to provide a wide range of facilities and potentials for conveying different types of context, accessibility, safety, diversity, and digestibility for all kinds of users with different mental or physical disorders.



Role of different kinds of learning methodologies such as micro learning [39], game-based learning [40], location-based learning [41], and ICT-based learning [42] in reinforcing the patients to overcome their barriers, should not be disregarded in this respect. In addition to the above-mentioned parameters, there also exist some software and GUI peculiarities which play a significant role in designing appropriate assisting applications in this domain. These may not only attract users, but also facilitate their exercises. In this paper, we would like to have a comparative study on these parameters within the existing applications.

III. SIGNIFICANT PECULIARITIES OF THE EXISTING SOFTWARES AS WELL AS THEIR GUI

Graphical User Interface (GUI) achieves a high role in facilitating communication between the system on the one side and the therapist or the patient on the other side. GUIs should therefore be designed in a way that such goals can be achieved most plausibly. To approach understanding what certain protocols should be considered in designing an optimal GUI, it should first be studied how the existing GUIs holding certain software peculiarities have in practice behaved toward the corresponding pathological disorders, and how these peculiarities can in turn be responsible for achieving the related success or failure. It should then be analyzed, based on the existing learning theories, how these considerations can be justified from epistemological viewpoint. What we have done in our approach, is to conduct such a study.

To conduct our comparison, we first need some significant peculiarities both for "software" and "GUI", based on which such entities can be characterized. To specify such peculiarities we first referred to some existing resources (including both accreditation websites and research papers) discussing the capabilities of the entities, and then out of the peculiarities mentioned for them, we selected those items which were believed to be of high significance for the area of Occupational Therapy through Mobile Applications [3, 4, 9, 13, 14, 15, 16, 17, 30, 31, 32, 33, 34, 35, 36, 37, 38, 43]. Within the process of selection, we paid mostly attention to those peculiarities which of least correlation in a thematic sense. We finally ended up with those elaborated in Table I.

Since GUI's peculiarities are hardly understandable, below we give a brief definition of them:

- **User in Control:** This master key has an undeniable role in the GUI efficiency, which mentions that users should have the authority of controlling the software behavior (for instance asking for users' confirmation for every vital command like "Quit", "Delete", and so on), as well as altering parts of interface in order to customize the GUI based on their own needs, skills, and habits.
- **Directness:** Users should be directed to their purpose through the shortest path, and by the minimal set of tasks in order to become able of retouching GUI objects in the most

direct manner.(allowing users to make shortcuts is the most common example of directness).

- **Consistency:** A GUI should be well-known and predictable for all sorts of users, who have been adapted to other software products so far. Therefore all facets of the GUI including labels, layout, behavior,... should be compatible with other software products' GUI in order to reduce the learning process of the software, and makes agile its adoption (objects dragging function in all standard smart phones is the same, according to which users should hold their finger on the object and move the object to their favorable place without removing their finger).

- **Forgiveness:** This peculiarity leads developers to make users' actions reversible and recoverable. Therefore users do not have to start from the first step once a mistake has occurred in the processes of choosing, typing, deleting, or even closing a window. The purpose of forgiveness is to allow users to become familiar with the software, and learn how to use it gradually (Recycle bin, where deleted objects are able to be restored is an example of "Delete" function forgiveness).

- **Feedback:** This maxim has a close relevance to the user in control peculiarity as though users should be provided with immediate visual response to their actions, which allows them to understand which processes are being executed (If the user has deleted an object, he/she should then immediately feel that the object does not exist anymore).

- **Aesthetics:** is proved to be the most discernible aspect of GUIs' design, since it is the first layer which is presented to users. Aesthetics peculiarities are proportion, symmetry, color, lines, texture, balance, flow, ..., which have a crucial role in attracting users, and can affect users' behavior and cognitive processes as well in visual facet; harmony is believed to depend on an appropriate arrangement of different fragments which leads finally to a pleasant perception by eyes (There have been lots of software products led to failure due to overlooking aesthetics aspects even though they have been rich in other aspects).

- **Simplicity:** A user interface should not be complicated to use and to learn either, because the more a GUI is designed simply and directly around what its users demand, the more success the system will gain. On the other hand, in order to optimize simplicity, developers should benefit from a type of prior knowledge about targeted users' level of education, age, gender, and purpose so as to provide the apt level of complexity (A comparison between Yahoo and Google represents the significance of simplicity in attracting users).

All kinds of mobile applications are somehow trying to teach a concept to the users. The concept can be even how to use the application. Therefore all of mobile applications should have been designed based on a learning theory. There are lots of learning theories which have been used in favor of different purposes, and in different styles such as "Behavioral Learning Theory", "Cognitive Learning Theory", "Information Processing Learning Theory" and "Constructive Learning Theory" [44, 45]. So learning theories should be elicited in

order to be recognized in applications. Here are the characteristics of some widespread learning theories, based on which mobile applications make sense.

IV. FACTS DERIVED THROUGH THE COMPARATIVE STUDY WITH RESPECT TO THE SOFTWARES USED

Looking at the content of Table I, the following facts seem to be consistent:

- "Reminders" with both mental and physical applications, are useful for the patients, who may forget to use the application, or for therapists who probably have more than one patient. Therefore smart schedules (or in other words reminders) as well as "consistency" are mandatory in GUI peculiarities, because these sorts of applications have been designed for long term usages, in a way that all features should be consistent in order to optimize the rehabilitation process.
- "Patients location records" have manifested only in some applications, for mental therapy. As the patients may get lost, location records are thus believed to be extremely helpful in these cases.
- "Exercise management" has been presented chiefly in physical and occupational therapy applications due to the intense requirement of exercises stability. Here "Consistency" has a promising role, because sporadic and unorganized exercises are harmful for the patient in the sense of decelerating the rehabilitation process.
- "Patients records" has mostly appeared in those applications, that try to produce a report of the rehabilitation or treatment process. This is to help therapists and doctors to find out how much the patient is in trying mode, what strengths and weaknesses he/she has, and whether the application is helpful or not. Therefore the application should lead the patient as directedly as possible in order to optimize this procedure. Due to this "Directness" has been considered effectively in GUI peculiarities.
- "Activity monitoring" is apparently similar to "Patient records", but the reason of separation lies in the variation of those activities done by the patients which cannot be considered as patient records. Particularly, in mental applications activity, monitoring is much more required than patient records, because in mental cases the treatment reports are usually produced implicitly when the patient is working with the application. The other point is that "Activity monitoring" is mandatory for applications that are based on "Constructive" learning theory, let say those which are game-based. These kinds of applications have mainly obtained a high grade of "Simplicity" in order to avoid any complication in the procedure of using the application for those patients who are not as intelligent as the normal people.
- "Audio instructions" and "Written instructions" have manifested mostly in physical and mental therapy applications in order to increase digestibility of

instructions given by the application. In the meantime, with regard to some mental cases, they have been presented in a musical way in order to tranquilize the patient. In this way, "atheistic" has been shown to be very effective for this purpose.

Besides the above facts we also noticed some other points with regard to relations between GUI's peculiarities and types of disability as follows:

- GUIs used for patients with mental disorders, and those having disability in hand and fingers, is required to be more "simple" mainly because of limitation in their ability to interact mentally with the surrounding environment.
- For patients with mental problems like Autism, "Aesthetic Aspects" are believed to be more important simply because such patients may need more attention compared to other patients.
- For patients with mental disorders, who suffer from visual tracking and insomnia, "Feedback" and "Forgiveness" aspects are necessary to promote for continuous practice.
- GUIs used for patients with physical disorders, especially with disability in hand area, should be "Consistent". That is because patients have got used to some software's facets and we should facilitate their reactions with minimum learning efforts.
- For patients with mental problems, who need to concentrate on writing skill and visual tracking, "Directness" and "User in control" are two significant aspects which can navigate the patient directly and personalize software for easy use.
- The most important aspects for physical disorders are "Consistency" and "Simplicity" which have to be considered in GUI design.

From the point of learning theory considerations, it is mentionable that, majority of software applications listed in Table I, control the progress trend of patient, by his/her behavior and actions. Thus, most of them have been constructed based on "Behaviorism". There however exist some applications mostly for mental problems which actually engage the patient and ask him/ her to write a text or draw a painting. In these cases, the applications are benefitted from "Constructivism" learning theory. The other kinds of learning theories have not been observed in the applications which have been investigated in Table I.

V. CONCLUSION

In the paper a comparative study was done on the existing GUIs for the patients with a variety of physical as well as mental disorders. Comparison was based on the peculiarities belonging to GUIs themselves and the softwares used for them to make them respond effectively. A tabular form was made in this regard to show the functionalities of the existing GUIs in association with the corresponding peculiarities. It was seen that, for mental disorders such as autism wherein patients are tremendously introvert and thus have slight feeling for interaction with their environment, GUIs holding peculiarities

such as "Aesthetics", "Feedback", or "Forgiveness" are suggested. While for disorders in writing skills and visual tracking, the patient prefers to make use of GUIs with peculiarities such as "Directness" and "User in control". Taking such points into account, one may gain the opportunity to build a hybrid strategy for designing GUIs that can function in case of patients with mixed disorders. This is quite interesting since many patients are in reality suffering from a combination of disorders each having his/her own complication with regard to the format of interaction and thereby the GUI to be used. Developing such a hybrid strategy can be regarded as a promising future work for therapeutic mobile applications.

REFERENCES

- [1] K. Bazargan Harandi (2011). Abstract Information Visualization in Interactive 3D Virtual Environments: Conceptualization and Usability Evaluation. Thèse de doctorat: Université de Genève, 2011, no. SES 747, Edition SES - Université de Genève, ISBN 978-2-88903-005-7.
- [2] H. E. Moore IV, O. Andlauer, N. Simon, E. Mignot, "Exploring medical diagnostic performance using interactive, multi-parameter sourced receiver operating characteristic scatter plots," *Computers in Biology and Medicine*, vol. 47, pp. 120-129, 1 April 2014.
- [3] Ž. Mijailović, D. Milićev, "Empirical analysis of GUI programming concerns," *International Journal of Human-Computer Studies*, vol. 72, issues 10–11, pp. 757-771, October–November 2014.
- [4] Y. B. Salman, H. I. Cheng, P. E. Patterson, "Icon and user interface design for emergency medical information systems: A case study," *International Journal of Medical Informatics*, vol. 81, issue 1, pp. 29-35, January 2012.
- [5] C.T. Asque, A.M. Day, S.D. Laycock, "Augmenting graphical user interfaces with haptic assistance for motion-impaired operators," *International Journal of Human-Computer Studies*, vol. 72, issues 10–11, pp. 689-703, October–November 2014.
- [6] J. M. Schraagen, F. Verhoeven, "Methods for studying medical device technology and practitioner cognition: The case of user-interface issues with infusion pumps," *Journal of Biomedical Informatics*, vol. 46, Issue 1, pp. 181-195, February 2013.
- [7] S. Mazzoleni, M. Munih, A. Toth, J. Cinkelj, M. Jurak, J. V. Vaerenbergh, G. Cavallo, P. Soda, P. Dario, E. Guglielmelli, "Whole-body isometric force/torque measurements for functional assessment in neuro-rehabilitation: User interface and data pre-processing techniques," *Computer Methods and Programs in Biomedicine*, vol. 110, issue 1, pp.27-37, April 2013.
- [8] D. J. Brown, D. McHugh, P. Standen, L. Evett, N. Shopland, S. Battersby, "Designing location-based learning experiences for people with intellectual disabilities and additional sensory impairments," *Computers & Education*, vol. 56, pp.11–20, 2011.
- [9] Microsoft Corporation Staff, *The Windows Interface Guidelines for Software Design*, Microsoft Press, 2 edition, January 1, 1994.
- [10] L. C. Mechling, and E. J. Savidge, "Using a personal digital assistant to increase completion of novel tasks and independent transitioning by students with autism spectrum disorder," *Journal of Autism Developmental Disorders*, vol. 39, no. 10, pp. 1420-1434, 2009.
- [11] T. Gentry, J. Wallace, C. Kvarfordt, and K. B. Lynch, "Personal digital assistants as cognitive aids for high school students with autism: results of a community-based trial," *Journal of Vocational Rehabilitation*, vol. 32, pp. 101–107, 2010.
- [12] J. Mintz, C. Branch, C. March, and S. Lerman, "Key factors mediating the use of a mobile technology tool designed to develop social and life skills in children with Autistic Spectrum Disorders," *Computers & Education*, vol. 58, pp. 53-62, 2012.
- [13] IoT Session, www.iot-session.com
- [14] Shelby's Quest, www.doodletherapyapps.com
- [15] Brain works, www.sensationalbrain.com
- [16] Autism Learning Games Camp Discovery, www.campdiscoveryforautism.com
- [17] Find me, www.interface3.com/findme
- [18] S. Stapleton, M. Adams, and L. Atterton, "A mobile phone as a memory aid for individuals with traumatic brain injury: A preliminary investigation," *Brain Injury*, vol. 21, no. 4, pp. 401–411, 2007.
- [19] T. K. Wade, and J. C. Troy, "Mobile phones as a new memory aid: A preliminary investigation using case studies," *Brain Injury*, vol. 15, no. 4, pp. 305–320, 2001.
- [20] T. Hart, K. Hawkey, and J. Whyte, "Use of a portable voice organizer to remember therapy goals in traumatic brain injury rehabilitation: A within subjects trial," *Journal of Head Trauma Rehabilitation*, vol. 17, no. 6, pp. 556–570, 2002.
- [21] H. J. Kim, D. T. Burke, M. M. Dowds, K. A. R. Boone, and G. J. Park, "Electronic memory aids for outpatient brain injury: Follow-up findings," *Brain Injury*, vol. 14, no. 2, pp. 187–196, 2000.
- [22] H. J. Kim, D. T. Burke, M. M. Dowds, and J. George, "Case study: Utility of a microcomputer as an external memory aid for a memory-impaired head injury patient during in-patient rehabilitation," *Brain Injury*, vol. 13, no. 2, pp. 147–150, 1999.
- [23] J. M. Fleming, D. Shum, J. Strong, and S. Lightbody, "Prospective memory rehabilitation for adults with traumatic brain injury: A compensatory training program me," *Brain Injury*, vol. 19, no. 1, pp. 1–13, 2005.
- [24] T. Gentry, "PDAs as cognitive aids for people with multiple sclerosis," *American Journal of Occupational Therapy*, vol. 62, no. 1, pp. 8–27, 2008.
- [25] T. Gentry, J. Wallace, C. Kvarfordt, and K. B. Lynch, "Personal digital assistants as cognitive aids for individuals with severe traumatic brain injury: A community-based trial," *Brain Injury*, vol. 22, no. 1, pp. 19–24, 2008.
- [26] G. M. Giles, and M. Shore, "The effectiveness of an electronic memory aid for a memory impaired adult of normal intelligence," *American Journal of Occupational Therapy*, vol. 43, no. 6, 1989, pp. 409–411.
- [27] B. A. Wilson, A. D. Baddeley, and J. M. Cockburn, "How do old dogs learn new tricks: Teaching a technological skill to brain injured people," *Cortex*, vol. 25, pp. 115–119, 1989.
- [28] B. Richards, L. Leach, and G. Proulx, "Memory rehabilitation in a patient with bilateral dorsomedial thalamic infarcts," *Journal of Clinical and Experimental Neuropsychology*, vol. 12, pp. 395, 1990.
- [29] Operation Reach Out, www.militaryfamily.com/downloads/apps/military-suicide-prevention-operation-reach-out/
- [30] BellyBio, www.bellybio.com
- [31] Optimism, www.findingoptimism.com
- [32] iSleep Easy, www.meditationoasis.com/smartphone-apps/iphone-application-support/isleep-easy-free-app-support
- [33] PT and OT helper, www.ptandot-helper.com
- [34] Dexteria VPP, www.dexteria.net
- [35] Occupational therapy, www.angelfire.com/me/xutum
- [36] Therapy Boss, www.pragmait.com
- [37] Physical Therapy Home Exercises, www.facebook.com/pages/PTGenie
- [38] Physical Therapy for Kids, <https://itunes.apple.com/us/app/physical-therapy-for-kids/id923005281?mt=8>
- [39] T. Hug, *Didactics of microlearning: Concepts, discourses and examples*. Munster. Waxmann Verlag, 2007.
- [40] M. A. Hersh, and B. Leporini, Accessibility and usability of educational games for disabled students. In C. Gonzalez (Ed.), *Student usability in educational software and games: Improving experiences*, IGI Global, 2012, pp. 1–40.
- [41] D. J. Brown, D. McHugh, P. Standen, L. Evett, N. Shopland, and S. Battersby, "Designing location-based learning experiences for people with intellectual disabilities and additional sensory impairments," *Computers & Education*, vol. 56, pp. 11–20, 2011.
- [42] M. Hersh, "Evaluation framework for ICT-based learning technologies for disabled people," *Computers & Education*, vol. 78, pp. 30–47, 2014.

ICIT 2015 The 7th International Conference on Information Technology

doi:10.15849/icit.2015.0104 © ICIT 2015 (<http://icit.zuj.edu.jo/ICIT15>)

- [43] M. Zen, "Metric-Based Evaluation of Graphical User Interfaces: Model, Method, and Software Support," The Fifth ACM SIGCHI Symposium on Engineering Interactive Computing Systems (eics2013), UK, June 24–27, 2013.
- [44] A. Saif, Modern Educational Psychology, Doran Pub., 2010.
- [45] P. Kadivar, Psychology of Learning, Samt Publication, 2010.

TABLE I. OCCUPATIONAL THERAPY APPLICATION'S PECULIARITIES

Name of Software for Occupational Therapy	Software Peculiarities										GUI Peculiarities						Type of Disability			Learning Theory Considerations		
	Reminders	Patient's Location Records	Exercise Management	Patient Records	Activity Monitoring	Audio Instructions	Written Instructions	Operation System		Deployment	Status	User in Control	Directness	Consistency	Forgiveness	Feedback	Aesthetics	Simplicity	Physical		Mental	Mento-Physical
								Desktop	Android										Web-based		Installed	
PT and OT helper: Fingers Hand ⁽¹⁾	✓		✓	✓		✓	✓	✓	✓		✓	M	M	H	L	M	M	L	Fingers and Hand			Behaviorism
PT and OT helper: Ankle ⁽²⁾	✓		✓	✓		✓	✓	✓	✓		✓	M	M	H	L	M	M	L	Ankle			Behaviorism
PT and OT helper: Elbow ⁽³⁾	✓		✓	✓		✓	✓	✓	✓		✓	M	M	H	L	M	M	L	Elbow (for tennis players and golf players in two versions)			Behaviorism
Dexterity VPP ⁽⁴⁾			✓	✓	✓			✓	✓	✓	✓	H	H	M	L	M	H	H	Fingers and Hand			Constructivism
iOT Session ⁽⁵⁾				✓				✓		✓		L	M	H	L	M	M	M	Visual Perception, scanning, Handwriting	Autism, developmental delays, learning disabilities, ADHD	✓	Behaviorism
Occupational Therapy ⁽⁶⁾						✓			✓	✓		L	L	H	L	L	M	M	Useful OT Instruments, designed for daily activities			Behaviorism and Constructivism
Therapy BOSS ⁽⁷⁾	✓		✓	✓			✓	✓	✓	✓		M	M	M	M	M	L	M	depends on the therapist (can be used for every parts)			Behaviorism
Ask the OT ⁽⁸⁾						✓	✓		✓	✓		M	L	L	L	L	M	H	Information hotspot about OT,			Behaviorism and Constructivism
Shelby's Quest ⁽⁹⁾			✓	✓	✓		✓		✓		✓	M	M	M	M	M	M	H	Visual Perceptual Deficits, finger isolation	Autism, developmental delays, fine motor delays, AS and DS	✓	Constructivism
Ready to Print ⁽¹⁰⁾			✓	✓				✓	✓	✓		M	M	M	L	M	M	H	Visual Perceptual, Finger isolation	Fine motor skills, pre-writing skills, developmental delays	✓	Constructivism
Brain Works ⁽¹¹⁾	✓	✓	✓	✓	✓			✓	✓	✓		M	H	H	M	H	M	H		sensory processing disorders, Autism, ADHD, ADD, learning disabilities		Behaviorism
Letter Reflex ⁽¹²⁾			✓	✓	✓			✓		✓		H	H	M	L	M	H	H		Writing skills, letter reversal & backwards recognition		Constructivism

Name of Software	Software Peculiarities	GUI Peculiarities	Type of Disability	Learning Theory
-------------------------	-------------------------------	--------------------------	---------------------------	------------------------

	Reminders	Patient's Location Records	Exercise Management	Patient Records	Activity Monitoring	Audio Instructions	Written Instructions	Operation System		Deployment		Status		User in Control	Directness	Consistency	Forgiveness	Feedback	Aesthetics	Simplicity	Physical	Mental	Mento- Physical	
								Desktop	IOS	Android	Web-based	Installed	Open- source								Proprietary	Type of Pathological Disorder		
Dexteria Dots ⁽¹³⁾			✓				✓	✓		✓		✓	H	H	M	H	H	H	H		visual tracking & fine motor skill, visual memory, motor planning		Constructivism	
Dexteria Dots 2 ⁽¹⁴⁾			✓	✓			✓	✓		✓		✓	H	H	M	L	M	H	H		visual tracking & fine motor skill, visual memory, motor planning		Behaviorism	
Physical Therapy Home exercises ⁽¹⁵⁾			✓	✓			✓	✓		✓	✓		M	M	M	M	M	M	H	Fingers, Hand, Feet, Elbows, Knees			Behaviorism	
Physical Therapy for Kids ⁽¹⁶⁾			✓				✓	✓		✓		✓	M	M	H	L	M	H	H	Hands, Arms, Feet, Knees, Elbows, Waist			Behaviorism	
My Health Lounge Physical Therapy ⁽¹⁷⁾	✓		✓	✓	✓	✓	✓	✓		✓		✓	M	M	M	L	M	H	H	Legs, Knees, Elbows, Waist, Spine			Behaviorism	
Autism Learning Games Camp Discovery ⁽¹⁸⁾				✓	✓		✓	✓		✓	✓		M	M	M	M	M	M	M		Autism		Behaviorism	
Find me (autism) ⁽¹⁹⁾				✓			✓	✓		✓	✓		M	L	M	M	M	H	H		Autism		Behaviorism	
BellyBio Interactive Breathing ⁽²⁰⁾		✓		✓				✓		✓	✓		M	H	M	L	M	M	H		Breathing and stress management		Behaviorism	
Operation Reach Out ⁽²¹⁾				✓		✓		✓	✓	✓		✓	M	L	L	M	M	H	H		Depression		Behaviorism	
Deep Sleep With Andrew Johnson ⁽²²⁾	✓					✓		✓		✓		✓	L	H	M	M	H	H	H		Insomnia		Behaviorism	
Optimism ⁽²³⁾	✓		✓	✓	✓		✓	✓		✓		✓	M	L	M	M	M	H	M		Depression and Bipolar		Behaviorism	
iSleep Easy ⁽²⁴⁾			✓		✓		✓	✓		✓		✓	L	H	M	M	M	M	H		Insomnia		Behaviorism	

www.ptandot-helper.com ^{(1) (2) (3)}

www.dexteria.net ^{(4) (12) (13) (14)}

www.iot-session.com ⁽⁵⁾

www.angelfire.com/me/xutm ⁽⁶⁾

www.pragmat.com ⁽⁷⁾

www.askdrcovington.com ⁽⁸⁾

www.doodletherapyapps.com ⁽⁹⁾

www.apps.essare.net/app/ready-to-print ⁽¹⁰⁾

www.sensationalbrain.com ⁽¹¹⁾

www.facebook.com/pages/PTGenie ⁽¹⁵⁾

www.dfwapps3.com ⁽¹⁶⁾

www.bluejayhealth.com ⁽¹⁷⁾

www.campdiscoveryforautism.com ⁽¹⁸⁾

www.interface3.com/findme ⁽¹⁹⁾

www.bellybio.com ⁽²⁰⁾

www.militaryfamily.com/downloads/apps/military-suicide-prevention-operation-reach-out/ ⁽²¹⁾

www.relaxationapps.com ⁽²²⁾

www.findingoptimism.com ⁽²³⁾

www.meditationoasis.com/smartphone-apps/iphone-application-support/isleep-easy-free-app-support ⁽²⁴⁾

New Technique of Forensic Analysis for Digital Cameras in Mobile Devices

Jocelin Rosales Corripio, Ana Lucila Sandoval Orozco, Luis Javier García Villalba

Group of Analysis, Security and Systems (GASS)
Department of Software Engineering and Artificial Intelligence (DISIA)
Faculty of Information Technology and Computer Science, Office 431
Universidad Complutense de Madrid (UCM)
Calle Profesor José García Santesmases, 9
Ciudad Universitaria, 28040 Madrid, Spain
Email: jocelinr@ucm.es, {[asandoval](mailto:asandoval@ucm.es), [javiervg](mailto:javiervg@ucm.es)}@fdi.ucm.es

Abstract— Nowadays, forensic analysis of digital images is especially important, given the high use of digital cameras in mobile devices. The identification of the device type or the make and model of image source are two important branches of forensic analysis of digital images. In this paper we have addressed both, with an approach based on different types of image features and the classification using support vector machines. The study mainly has focused on images created with mobile devices and as a result, the techniques and features have been adapted or created for this purpose. There have been a total of 36 experiments classified into 5 sets, in order to test different configurations of the techniques. In the configuration of the experiments were taken into account among other things the future use of the technique by the forensic analyst in real situations and creating experiments with high technical requirements.

Keywords— *Forensics Analysis, digital image, image source acquisition identification, image noise features, image color features, image quality metrics, image wavelet features*

I. INTRODUCTION

Currently, the demand for mobile devices (mobile phones, smartphones, tablets, etc.) increases year by year despite the global economic crisis. According to Gartner [1] in 2013 smartphone sales grew 42.3% over the previous year, outnumber for the first time the sales of feature phones. We must not overlook the emergence in today's society of such devices in our day to day life. Increasing storage capacity, usability, portability and affordability, have allowed mobile devices to be present in several activities, places and events of daily life. A consequence of its widespread use, is that digital images can be used as silent witnesses in judicial proceedings (child pornography, industrial espionage, ...), and in many cases crucial pieces of an evidence of a crime [2].

Forensic analysis of digital images can be mainly divided into two branches [3]: tamper detection and image source identification. This work focuses on the first branch. Also, since mobile device cameras have some characteristics that make them different from the rest, this work focuses on images from this type of devices. In this paper, we propose a method to image source acquisition in mobile devices. The objective of this approach is to identify make and model from a group the different images into disjoint sets in which all their images belong to the same device. This paper is structured into 5 chapters, being the first this introduction. The rest of the paper is structured as follows. Section 2 shows carries out a state of the art of techniques and algorithms for identifying the source

type and source acquisition identification. Section 3 shows different sets of features (Noise, Color, Image Quality Metrics (IQM) and Wavelets) used by the algorithms and techniques of forensic analysis. In section 4, a set of experiments for the identification of device type and the source acquisition identification of the image are performed. In these experiments we use the set of the features previously presented and the algorithms of the techniques. Finally, section 5 shows the main conclusions of this work and some future work lines.

II. RELATED WORK

The main techniques of digital image forensics for identifying the source of image acquisition and the main work of the analysis. The success of these techniques depends on the assumption that all the images acquired by the same device have intrinsic features. The features which are used to identify the make and model of a digital camera are derived from the differences between the techniques of image processing technologies and the components which are used. The biggest problem with this approach is that different models of digital cameras use components of a small number of manufacturers, and the algorithms used are also very similar between models of the same brand. According to [4] for this purpose four groups of techniques can be established depending on their base: lens system aberrations, Color Filter Array (CFA) interpolation, image characteristics, and sensor imperfections.

Techniques Based on Image Features use a set of features extracted from the content of the image to identify the source. These features are divided into three groups: color features, Image Quality Metrics (IQM) and wavelet domain statistics.

[5] proposes a method to identify the source using the following features: color features, image quality metrics and frequency domain. The study adopted the wavelet transforms as a method to calculate the wavelet domain statistics and use a Support Vector Machine (SVM) for classification. In experiments digital cameras and mobile devices were used. The results obtained in different experiments show results between 61.7% and 99.72% accuracy.

In [6] authors extend the source identification to different devices such as mobiles, phones, digital cameras, scanners and computers. In this proposal they base it on the differences in the image acquisition process to create two features groups: color interpolation coefficients and noise features. In the experiments they use five smartphone models, five digital camera models and four scanner models to identify the source type. Their experiments showed an overall result of 93.75% accuracy. Identifying the maker and model of five mobile phone models resulted in an accuracy of 97.7%.

In [7] a method for source camera identification is proposed through the extraction and classification of wavelet statistical features. Finally 216 first-order wavelet features and 135 second order co-occurrence features is obtained. The most representative features are selected using an Sequential Forward Featured Selection (SFFS) algorithm and they are classified using a SVM. Identification success average of 98% the set of all cameras and an average success rate of 96.9% for the three cameras of the same model is achieved.

[13] performs experiments with common imaging features to identify the source: wavelet, color, IQM, statistical features of difference images and statistical features of prediction errors. In the experiments, different combinations of different types of features are used and a SVM for classification of different devices. Ten different cameras from four different makers with 300 images from each camera (150 for training and 150 for testing) and a resolution of 1024x1024 is used. Using all the features a score of 92% success rate is obtained. Moreover experiments were performed to check the robustness against three of the most common alterations in digital images: JPEG compression, cropping and scaling.

In [9] a technique for image source identification is proposed using ridgelets and contourlets subbands statistical models. After the feature extraction a SFFS algorithm is used for feature election and a SVM for classification. The method based on 216 wavelet features is considered useful only for the representation of a dimension, the approach based on ridgelets uses 48 features, and the approach based on contourlets includes a total of 768 features. In experiments with three cameras from different makers success rates are between 99.5% and 99.8%.

In [10] a method using the marginal density Discrete Cosine Transform (DCT) coefficients in low-frequency coordinates and neighboring joint density features from the

DCT domain is proposed. Furthermore, hierarchical clustering and SVM is used to detect the source of acquisition of the images. In experiments with images from five smartphone models of four makers an accuracy of between 86.36% and 99.91% was obtained, achieving the best results with a linear SVM kernel.

I. PROPOSED WORK

Regarding classification, in [11] a study of different classification methods such as distance-based classifiers, Bayesian classifiers, neural networks, clustering algorithms and SVM classifiers is performed. As can be observed in the review, the use of SVM classifiers is widely used for these purposes. The kernel choice depends, among other factors, on the nature of the data to be classified. This paper will use an SVM classifier with Non-linear RBF kernel, as it is recommended for use when there is no a priori information about the data. The parameters for the SVM are the same as those used in [12]. Likewise, the option chosen is the most widely used one by the most recent precise works and they present good results. There are many implementations of SVM classifiers; particularly in this work we opted to use the LibSVM library [13].

The set of features to be used can be classified into four major groups, depending on the nature of their obtaining: noise features (16 features), color features (12 features), IQM (40 features) and wavelets (81 features). A detailed analysis on each of the aforementioned feature sets will be performed below.

A. Noise Features

One of the objectives is to get a set of features that allow us to differentiate between the different types of devices. To do this we firstly take into account that digital cameras use a two-dimensional array sensor whereas most scanners use a linear array sensor. In the case of scanners, the linear arrangement of the sensor moves to generate the entire image, so it is expected to find the periodicity of the sensor noise within the rows of the scanned image. On the other hand, there is no reason to find sensor noise periodicity within the columns of the scanned image. In the case of digital cameras this type of noise periodicity does not exist. This difference can be used as a basis to discriminate between different types of devices. Noise features extraction is based on [14].

Let I an image of $M \times N$ pixels, M as the rows and N as the columns. We denote I_{noise} the noise of the original image and $I_{denoise}$ is the image without noise.

$$I_{noise} = I - I_{denoise} \quad (1)$$

Then, each color component of the image without noise is subtracted to each color component of the original image, with which we obtain noise components of each pixel disaggregated for each color component.

The image original noise I_{noise} can be modeled as the sum of two components, the constant noise $I_{noiseconstant}$ and random noise $I_{noiserandom}$. For scanners constant noise only depends of the column index, because the same sensor is

moved vertically to generate the complete image. The average noise of all columns can be used as a pattern reference $I_{noiseconstant}(1, j)$ because the random noise components were cancelled. For detecting the similarity between different rows with the pattern reference, we use the correlation of these rows with the pattern.

$$corr(X, Y) = \frac{(x-\bar{x}) \cdot (y-\bar{y})}{\|x-\bar{x}\| \cdot \|y-\bar{y}\|} \quad (2)$$

Then the same process is performed to detect the similarity of the columns with the pattern reference. After obtaining the correlation between rows and between columns we will go to obtain the feature set. It should be noted at the time of obtaining the features, that in the case of scanners the orientation of the image is critical, because features obtained will be completely different.

For each type of correlation first order statistical values are obtained, which are: mean, median, maximum and minimum. Also, the ratio features between rows and columns correlations are added. Finally the average noise per pixel feature was included. This feature does not depend on rows or columns correlations with the reference pattern, but is independent and it can distinguish between different types of devices, such as computer generated images. In total a set of 16 features are obtained: 7 rows features, 7 columns features, the ratio between rows and columns correlations and the average noise per pixel.

B. Color Features

The configuration of the CFA filters, the demosaicing algorithm and color processing techniques mean that signals in the color bands may contain treatments and specific patterns. In order to determine the differences in color features for different camera models, it is necessary to examine the first and second order statistics of the pictures taken with them.

- *Pixels average value*: This measure is performed for each RGB channels (3 features).
- *Correlation pair between RGB bands*: This measure expresses the fact that depending on the structure of the camera, the correlation between the different color bands can change (3 features which come from measuring the correlation between the RG, RB and GB bands).
- *Neighbor distribution center of mass for each color band*: This measure is calculated for each band separately (3 features). Firstly, the total number of pixels for each color value is calculated, obtaining a vector with 256 components. Then, with these calculated values the sum of neighboring values are obtained.
- *Energy ratios between pairs RGB*: This feature depends on the white dots correction process of the camera (3 features)

C. Image Quality Metrics

Different camera models produce images of different quality. There may be differences in image brightness, sharpness or quality color. These differences propose a set of quality metrics features that help us to distinguish the image source. There are different IQM categories: measures based on the pixels differences, measures based on correlation and measures based on spectral distance. For obtaining this set of

metrics, a filtered image in which the noise of the original image is reduced to perform different calculations is needed in addition to the original image. For this, a Gaussian filter that allows us to perform image smoothing is used. After the core is obtained, it is normalized, so that the sum of all its components is 1. This is necessary to obtain a smooth image but with the same colors as the original. The normalization is performed dividing each component by the sum of the values of all the components. For obtaining the metrics a filter with a 3x3 kernel with $\gamma = 0.5$ is used. Following the specification of the 40 IQM features based on [8].

- *Czekonowsky distance*: The Czekonowsky distance is a useful metric for comparing vectors with no negative components as in the case of color images.
- *Minkowsky metrics*: Minkowsky metrics for $\gamma = 1$ and $\gamma = 2$.
- *Normalized Cross Correlation*: The closeness between two digital images can also be quantified in terms of a correlation function. The quality metric of the normalized cross-correlation measurement for each image band k .
- *Structural Content*: The structural content of an image quality metric is defined for each band k .
- *Spectral Measures*: The Discrete Fourier Transform (DFT) of the original image and the smoothed image, denoted as $\tau_k(u, v)$ and $\hat{\tau}_k(u, v)$ for a band k .
- *Measures based on the human visual system*: Images can be processed by filters which simulate the Human Visual System (HVS). One of the models used for this is a band-pass filter with a transference function in polar coordinates.

D. Wavelet Features

Due to the deterministic property of the sensor pattern noise which is present in an image, this pattern can be used as a footprint to identify the device that generated the image under investigation. It can be said that the sensor pattern noise is to a digital camera as a fingerprint is to a human being. To identify the acquisition source we require an algorithm that allows us to extract the sensor noise and another that allows us to obtain the features of the fingerprints obtained in order to classify and identify them.

Taking the main ideas from [15] as a reference, algorithm 1 is proposed to extract sensor noise.

Algorithm 1: Extracting PRNU

1. Apply a wavelet decomposition in 4 levels to I ;
2. **ForEach** wavelet decomposition level **do**
3. **ForEach** component $c \in \{H, V, D\}$ **do**
4. Compute the local variance;
5. **If** (adaptive variance)
6. Compute 4 variances with windows of size: 3, 5, 7 and 9 respectively;
7. Select the minimum variance;
8. **else**
9. Compute the variance with a window of size 3;
10. Compute noiseless wavelet components applying the Wiener filter to the variance;

11. Obtain I_{clean} by applying the inverse wavelet transform with clean components calculated;
12. Obtain the sensor noise with $I_{noise}=I- I_{clean}$;
13. Apply zero-meaning to I_{noise} ;
14. Increase the green channel weight with

$$I_{noise} = 0.3 \cdot I_{noise_R} + 0.6 \cdot I_{noise_G} + 0.3 \cdot I_{noise_B};$$

Finally, a total of 81 features (3 channels x 3 wavelet components x 9 central moments) are calculated using algorithm 2.

Algorithm 2: Extracting features

1. Separate R, G and B color channels of I_{noise} ;
2. **ForEach** color channel **do**
3. Apply a wavelet decomposition in 1 level;
4. **ForEach** component $c \in \{H,V,D\}$ **do**
5. Compute k central moments with

$$m_k = \frac{1}{n} \sum_{i=1}^n |c - \bar{c}|^k$$

II. EXPERIMENTS AND RESULTS

We performed the classification of images on closed set of elements, i.e., the classes of the elements used in training are the same classes as those used in the test. The images used in the training stage are not used in the testing stage.

In order to evaluate the source device type identification we will use an image set composed of: images from mobile phones, images obtained from a scanner, and a computer-generated images. 200 images are used from each set, 100 for the SVM training and 100 for testing. All images have a resolution higher than 1024x768. There is no restriction on the content of the image or the camera configuration parameters at the time of the acquisition.

Images from 7 smartphones: iPhone 4s (I1), Blackberry 8520 (BB), Huawei U8815 (HU), LG P760 (LG2), Nokia 800 (N1), Samsung GT-I9001 (S1) and Sony C2105 (SE1). For images from scanners and computer-generated images, our own sources and the Flickr website were used. As a second filter for scanned images, those which had the tag “scanned images” and made reference to a retail scanner model were used. For the experiments we have taken into account the following configuration parameters: size of crop applied to the image, crop position (centered or upper-left corner) and application of different feature sets (Noise Features, Color Features, IQM Features and Wavelet Features).

Table I shows the results of success rates to evaluate the source device type identification between Camera (A), Computer (B) and Scanner (C), and the configuration parameters used in the 10 experiments.

From the analysis of the results, general and specific conclusions about the various configurations used in each

experiment can be obtained. Encompassing all the experiments, it is observed that success rates are not excessively high (60.42% on average and 71.30% in the best case); it can be concluded that this technique is not particularly suitable for this purpose. It is important to emphasize, as noted above, that the number of different makes and models used for this experiment is high, which predictably causes success rates to drop. That being said, it should be noted that this study does provide interesting results on the configuration parameters used, since between the best and the worst result there is a difference in the average success rate of 23.48%.

TABLE I. TPR WITH EQUAL NUMBER OF DEVICES THAN CLUSTERS

Features	Crop Size	Crop Align	Device (%)			Average (%)
			A	B	C	
Noise	Full Size	-	70	54	57	59.95
	1024x768	Center	66	80	46	62.39
	800x600		76	60	49	60.68
	640x480		62	61	48	56.62
	1024x768	Upper-left corner	76	59	40	56.40
	800x600		65	38	44	47.72
640x480	74		54	37	52.88	
All Features	1024x768	Center	66	73	72	70.26
	800x600		69	74	71	71.30
	640x480		77	73	63	70.75
Average			69.9	61.3	51.4	60.42

Given the importance of mobile images today, below we will show the experiment performed to identify the acquisition source of images from mobile devices, i.e., the classification of an image set according to the make and model of the camera that generated them.

The results improve significantly when all the features to identify the source type are used. Given the high number of classes, the results can be qualified as acceptable, since the average success rate for all experiments carried out using these features is 70.77%. The experiments have been grouped into 3 groups with the aim of obtaining conclusions on: the use of different feature sets, crop size, the number of devices used for the classification, and the use of devices from the same manufacturer.

Table II shows the experiments in which 7 models of mobile devices from different manufacturers are used. Different types of combinations of features sets were tested. Most experiments were performed with a crop size of 1024x768, since as this is considered a large enough size to obtain good results, as shown in the previous experiments.

TABLE II. TPR WITH EQUAL NUMBER OF DEVICES THAN CLUSTERS

Features	Crop Size	Crop Align	I1	HU	LG2	N1	BB	S1	SE1	Average
All Features (Daubechies 8-tap)	1024x768	Center	93	96	80	94	91	70	85	86.54

Noise	1024x768	Center	41	42	35	18	40	40	62	37.67
Color	1024x768	Center	24	37	20	40	31	19	44	29.27
IQM	1024x768	Center	13	88	46	89	7	34	2	21.65
Wavelet Daubechies 8-tap	1024x768	Center	95	96	96	94	92	76	93	91.46
Wavelet Haar	1024x768	Center	95	87	97	70	86	56	91	81.84
Color + IQM + Wavelet Daubechies 8-tap	1024x768	Center	93	94	90	90	90	53	85	83.67
All Features (Daubechies 8-tap)	800x600	Center	91	96	84	92	95	56	85	84.41
All Features (Daubechies 8-tap)	640x480	Center	90	95	84	89	88	51	88	82.15

The experiment reveals that noise, color and IQM feature sets are individually completely invalid, since the best result obtains an 37.67% average success rate, which is unacceptable. With the remaining set of features (wavelets), two experiments were conducted using different types of wavelet: Daubechies 8-tap and Haar. The results show that Daubechies 8-tap obtains better results than Haar and the best results of all experiments (91.46%).

With respect to the different feature combinations, it is observed that when we use all the features good results are obtained (86.54% in the best case), since, although they are slightly worse than the best result, the difference is not very significant (4.92%). Also, the success rate when all the features are used subtly drops the smaller the crop size gets.

The combination of all the features except noise features, which are mainly focused on identifying the source type, yields an average success rate of 83.67%. These results, even if not bad, are far from those obtained with the wavelets and worse than when the combination of all features is used.

CONCLUSIONS

In this work we have presented various techniques for identifying mobile device images with respect to scanned and computer-generated images. Besides, other techniques that allow us to distinguish the acquisition source of smartphone images are presented. The techniques are based on the use of four feature sets (Noise, Color, IQM and Wavelets), on which adjustments have been made in order to improve the results for this specific type of devices. There have been experiments with the combination of the different feature sets, different crop sizes and positions, and wavelet functions. With regard to source type identification, the first general conclusion is that Noise features are discarded as invalid when the number of types of devices is greater than 2. In the experiments that used whole images and different crop sizes and positions, unacceptable results were obtained for identifying three types of devices (scanner, smartphone and computer). As discussed in the experiments, for these three types of devices there are dozens of different manufacturers and models, hampering classification. As a counterpart, forensic analysts may consider the application of the technique with Noise features for identifying the source type of images from mobile devices with respect to images from scanners and computers. The results are quite good at identifying the type when discerning between scanners and smartphones. The use of all the features significantly improves results, but as a general conclusion they are not good enough to be used in a serious situation. When

identifying the acquisition source of mobile device images, the results are much more encouraging. In all sets of experiments performed, there is at least one configuration that yields good results, always putting them into the context of the level of demand on this technique (a large number of devices or many devices from the same manufacturer).

ACKNOWLEDGMENT

The research leading to these results has been partially funded by the European Union's H2020 Program under the project SELFNET (671672). Part of the computations of this work was performed in EOLO, the HPC of Climate Change of the International Campus of Excellence of Moncloa, funded by MECED and MICINN. This work was supported by the "Programa de Financiación de Grupos de Investigación UCM validados de la Universidad Complutense de Madrid – Banco Santander".

REFERENCES

- [1] Gartner Says Smartphone Sales Grew 46.5 Percent in Second Quarter of 2013 and Exceeded Feature Phone Sales for First Time (2013). URL <http://www.gartner.com/newsroom/id/2665715>
- [2] M. Al-Zarouni, "Mobile Handset Forensic Evidence: a Challenge for Law Enforcement", in *Proceedings of the 4th Australian Digital Forensics Conference*, School of Computer and Information Science, Edith Cowan University, 2006.
- [3] T. Gloe, M. Kirchner, A. Winkler, R. Bohme, "Can We Trust Digital Image Forensics?", in *Proceedings of the 15th International Conference on Multimedia*, ACM Press, 2007, pp. 78–86.
- [4] T. Van Lanh, K. S. Chong, S. Emmanuel, M. S. Kankanhalli, "A Survey on Digital Camera Image Forensic Methods", *IEEE International Conference on Multimedia and Expo*, IEEE, 2007, pp. 16–19.
- [5] M. J. Tsai, C. L. Lai, J. Liu, "Camera/Mobile Phone Source Identification for Digital Forensics", in *Proceedings of the International Conference on Acoustics Speech and Signal Processing*, IEEE, 2007, pp. II-221–224.
- [6] C. Mckay, A. Swaminathan, H. Gou, M. Wu, "Image Acquisition Forensics: Forensic Analysis to Identify Imaging Source", in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2008, pp. 1657–1660.
- [7] B. Wang, Y. Guo, X. Kong, F. Meng, "Source Camera Identification Forensics Based on Wavelet Features", in *Proceedings of the International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, IEEE Computer Society, 2009, vol. 0, pp. 702–705.
- [8] Y. Hu, C. T. Li, C. Zhou, "Selecting Forensic Features for Robust Source Camera Identification", *Computer Symposium (ICS)*, 2010 International, 2010, pp. 506–511.
- [9] L. Ozparlak, I. Avcibas, "Differentiating Between Images Using Wavelet-Based Transforms: A Comparative Study", *IEEE Transactions on Information Forensics and Security*, IEEE, 2011, vol. 6, no. 4, pp. 1418–1431.

- [10] Q. Liu, X. Li, L. Chen, H. Cho, A. P. Cooper, Z. Chen, M. Qiao, A. H. Sung, "Identification of Smartphone-Image Source and Manipulation", *Advanced Research in Applied Artificial Intelligence, Lecture Notes in Computer Science*, Springer Berlin Heidelberg, Dalian, China, 2012, vol. 7345, pp. 262–271.
- [11] D. Michie, D. J. Spiegelhalter, C. C. Taylor, "Machine Learning, Neural and Statistical Classification", *Ellis Horwood*, 1994.
- [12] J. Rosales Corripio, D. M. Arenas González, A. L. Sandoval Orozco, L. J. García Villalba, J. C. Hernandez-Castro, S. J. Gibson, "Source Smartphone Identification Using Sensor Pattern Noise and Wavelet Transform", In *Proceedings of the 5th International Conference on Imaging for Crime Detection and Prevention (ICDP 2013)*, pp. 1–6 2013.
- [13] C. C. Chang, C. J. Lin, "LIBSVM: A Library for Support Vector Machines". Version 3.17, Abril 26, 2013. URL <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>
- [14] N. Khanna, A. K. Mikkilineni, E. J. Delp, "Scanner Identification Using Feature-based Processing and Analysis", *IEEE Transactions on Information Forensics and Security*, IEEE, 2009, vol. 4, no. 1, pp. 123–139.
- [15] J. Lukas, J. Fridrich, M. Goljan, "Digital Camera Identification from Sensor Pattern Noise", *IEEE Transactions on Information Forensics and Security*, IEEE, 2006, vol. 1, no. 2, pp. 205–214.,

Virtual Tourism Application through 3D Walkthrough: Flor De La Mar

Mohd Rahmat Bin Mohd Noordin

Faculty of Computer & Mathematical Sciences
MARA University of Technology
Jasin, Melaka, Malaysia
mrahmat.noordin@melaka.uitm.edu.my

Ismassabah Binti Ismail

Faculty of Computer & Mathematical Sciences
MARA University of Technology
Jasin, Melaka, Malaysia
isma@tmsk.uitm.edu.my

Muhammad Nur Aiman Bin Mohd Yahya

Faculty of Computer & Mathematical Sciences
MARA University of Technology
Jasin, Melaka, Malaysia
aiman_yahaya@gmail.com

Abstract— Tourism industry with well develop virtual tour is important to promote a country iconic places of interest such as Malacca. It is essential for Malacca as part of the tourism attraction to preserve the historical value for the tourist and also for the younger generation to learn from the history. Maritime Museum is one of the must-go-see places where it is built based on a Portuguese's ship that known as Flor de La Mar (FdLM). However, based on the site visit and the peer review that has been done, the existing virtual museum in Maritime Museum Malacca does not reflect the historical identity of the ship. In this paper, we will focus on development of FdLM 3D walkthrough application that would enhance the existing virtual museum with virtual and augmented reality implementation. Results from the evaluation have shown positive feedback where users are able to imagine the FdLM in real life and they prefer FdLM 3D Walkthrough compared to the existing virtual museum. Thus FdLM 3D Walkthrough application shall be adapted in the future to the Maritime Museum Malacca as tourist attraction.

Keywords— *virtual tourism, virtual reality, augmented reality, Malacca, Flor de La Mar (FdLM)*

I. INTRODUCTION

The Malacca Maritime Museum was designed as a replica of the original architecture of a Portuguese ship; Flor de La Mar (FdLM) which sank off in the coast of Malacca. The replica was constructed in 1990 and officially opens to public on June 13, 1994 [1]. It is vital to preserve this replica and its legendary historical identity in a new technological way to ensure that today and next generation keep their interest towards the ship and the history behind. This will help the Malacca State in establishing their famous Historical State title, along with the current technology interest.

Currently, there is an existing virtual museum for the FdLM in the official website. However, the current virtual museum

does not give the visitor the actual look and feel of FdLM and it looks nothing more than slide show. There are not enough multimedia elements that allow the visitor to interact with the virtual museum and provided only text with slide show. This is very disappointing as it limit the visitors' interaction due to the lack of multimedia elements. In addition, the slide show of FdLM does not help visitors to know the interior design of the ship. In term of design, this will help them on what to expect, even before they go to the Maritime Museum. Based on the initial investigation, FdLM is in desperate need of improvement.

Furthermore, there is not enough information about the ship from the virtual tour that limits the visitors' understanding. Historical information about the ship is very important to make

sure the visitors know every details of the ship. Implementing learning theory in the walkthrough will help solving this problem.

A 3D walkthrough application would be a solution to the current problem. The main purpose of FdLM 3D Walkthrough is to allow the tourist to experience the ship in a virtual environment and at the same time conveying brief history of the ship and Malacca in particular. With the help of virtual walkthrough, visitors should be able to imagine more clearly about the ship. Users are able to experience the ship in a new 3D environment where they are able to 'walk' inside the ship in a better visual. Besides, they are also able to emerge in the virtual environment and interact with the 3D objects. Users also have a better understanding and exposure about the ship from seeing a physical object in a solid form.

It is also important for the users to experience the ship in a real world environment. Users are able to have total immersion when the 3D ship overlay with their physical world with the implementation of augmented reality (AR). Engagement of FdLM 3D ship and the real world would help in maximizing the opportunity of interaction. The implementation of VR and AR for FdLM 3D Walkthrough offers a richer user experience. Information delivery of the historical side of the ship is much more entertaining and accurate. This application would also help in guiding the tourist.

II. RELATED WORKS

A. Virtual Reality

Oxford Dictionary defines the word virtual as almost or nearly as describe, but not completely or according to strict definition [2]. Meanwhile, the word reality is defined as a thing that is actually experience of or seen.

The word virtual reality means a vision of reality without its physical, or whatever else that is needed to build the reality [3]. VR is also a way for human to visualize, manipulate and interact with computers and extremely complex data [4]. It has been widely use in various fields, especially in medical, military and engineering. It is strongly suggests that VR is not only a medium or high end user interface, but it is an application which involves solution to real problems.

VR is an application that able to solve a particular problem with a well performed simulation that depends very much on the third 'I'; human imagination. The imaginative aspect of VR also refers to the human mind capacity to perceive non-existent things [5].

According to The Virtual Technology, Volume 1, it is also stated that VR is an integrated trio of immersion-interaction-imagination as illustrated in Fig. 1.

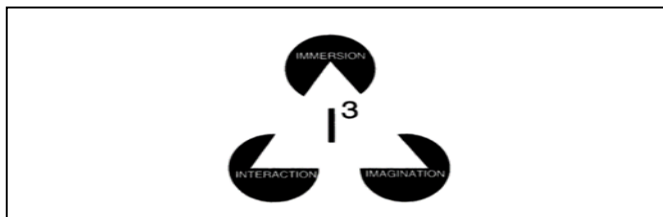


Fig. 1. The three I's of VR

The three pillars of VR would help the users to experience the look and feel of the virtual environment, thus contributes in their understanding of the historical ship.

B. Virtual Tourism in Malaysia

Information and communication technologies (ICT) in Malaysia has transformed drastically towards the tourism industry. This enhancement has contributed to a new paradigm of tourism industry to be one of the major evolving industries here. [6].

Tourism organizations are promoting and distributing their product and services with a lower cost internet. Internet has made it possible to interact in one-on-one marketing, other than being a basis for mass marketing [7].

Along with the globalization of tourism industry, accurate information is needed so it would be convenient for the tourist around the world to easily search needed information [6]. Hence, the delivery of information with 3D virtual reality and augmented reality technology would help in making that step much more imaginable and interactive for the tourist.

Virtual tourism is a great marketing tool to promote Malaysian uniqueness to the world. In fact it is an emerging tourism tool since the launch of Virtual Malaysia in 1997, an official virtual reality website to promote the country tourism. The Chief Executive Officer, Rohizam Muhammad Yusoff, highlights the organization's goal which is to utilize the VR technology by providing virtual tours destination, service and even promoting Malaysian products [8].

From the website, viewers are able to experience a 360-degree panoramic view of many Malaysia's places of interest such as the Tempurung Cave, a tourist attraction highlighted in the State of Perak. Furthermore, the virtual tour offers the closest possible sensory experience of any tourist destination, services, and product. Besides Virtual Malaysia, Islamic Art Museum Malaysia (IAMM) is also taking a step further as compare to other local museums by offering a virtual tour which also consist of 360-degree panoramic view

However, according to a journalist Chan Chun Yew [9], the Malaysian virtual tour is rather disappointing. The writer emphasizes disappointment when it was expected to be a 3D-movement in virtual world-ala Myst but somehow it is not like how he imagined. A similar issue occurred for the existing FdLM virtual, where there are only transitioning images.

Despite the increasing in ICT awareness and public multimedia literacy, Malaysia technology needs to improve vividly in order to go parallel with other developing countries. Multimedia application with combination of 3D walkthrough virtual reality and augmented reality could be the answer in dealing with this issue.

C. Augmented Reality (AR)

Different from virtual reality where a virtual environment replaces the physical world, AR superimposes the real world with additional information [10]. AR also helps to improve a

user perception and interaction with the physical world by supplementing with 3D virtual objects that immerse within the same space [11]. In this study, the 3D model of the ship is immersing in the real world through a marker-based detection.

There are three requirements of an AR application [12]. Combination of virtual elements such as 3D objects and the real environment is the basic requirement of AR application. According to Milgram and Kishino [13], it would be a helpful perspective to perceive AR as part of a reality-virtually continuum concept where it acts as a scale of ranging an environment to be completely real, and the other part is completely virtual. AR resides between these two ranging as shown in Fig. 2.

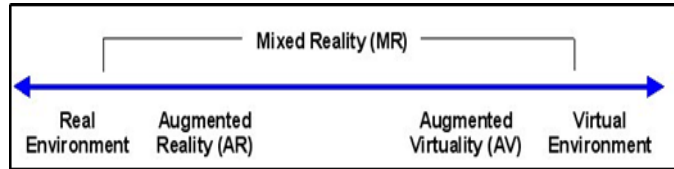


Fig. 2. Reality-virtually continuum

The second requirement of AR is the alignment of virtual elements with the real world. These means that the 3D object should have aligned and match together with the real surroundings.

Third, AR application should be in real-time interactivity to ensure that the virtual elements behave like a real element concurrent with the real space. However, this may depends on the changes in the perspective of the user, lighting conditions, occlusion and other physical laws [12].

A marked is needed to display the virtual elements as shown in Fig. 2. The purpose of this is to identify and position the 3D objects parallel with the real environment. When the registered marker is detected, the system will position the 3D coordinates of the marker pattern and place the virtual object on it [14]

Together with combination of 3D modelling and marker-based AR, visitors would be able to experience and interact with the FdLM ship as it merge with a reality perspective through a desktop camera.

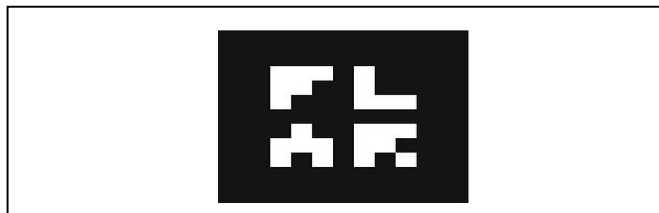


Fig. 3. Sample of AR Marker

Along with the help of a marker design for this application, the process of AR happens when the image is captured from the desktop camera. As shown in Fig. 4, a line detection method is used to find the marker edges to locate the intersection so that it will be easier to identify the marker's corner points [15].

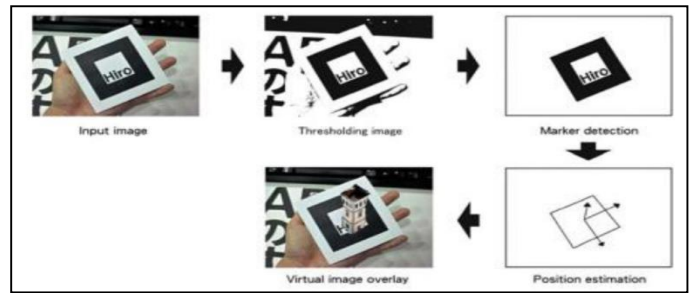


Fig. 4. Marker-based AR operates

Pattern with four corner is the most easy and accurate marker for AR process. Thus, a square based shape is the most suitable marker to be used [16].

III. METHODOLOGY

A step by step process of developing the FdLM 3D Walkthrough is crucial in order to develop an application that serves its objective and solve the current problem. Listed are the steps taken. (a) Initial Investigation, (b) Peer review, (c) Design and development, (d) Implementation of FdLM 3D Walkthrough and (e) Testing.

A. Initial Investigation

This process will help in finding specific details of FdLM in order to develop a 3D walkthrough and to prevent any misinformation with the final product. Thus, initial investigation is carried out to collect the information needed. Site visit is conducted to explore the information needed to develop the application.

Site visit is one of a crucial step to have an idea or brief information about the project. According to Princeton University [17], site visit can be defined as a visit under authorized ability to examine a site to focus its suitability for some endeavor. Undergo this process helped a lot in figuring out how FdLM looks like in real life as well as imagining how the historical ship should be designed.

Based on the initial investigation during the site visit, few photos and notes are taken down in order to collect information regarding the history of the ship. These are some of the information that has been collected:

- FdLM is made up of three levels
- The captain room is at the top level.
- The other two levels include artifacts, documents, and paintings from Malacca golden era.
- The replica ship is standing 34 meters high, and 8 meters width.
- The museum highlights Malacca as the Emporium of The East.

B. Peer Review

A peer review is done to identify the problem with the current FdLM virtual tour as a data gathering to find solutions needed to be implement in order to its virtual elements.

Peer review is a documented review done of associates or peers, who is having skill in the topic to be reviewed or a subset of the topic to be reviewed at any rate proportion which at least equivalent to that need for original work and also independent for the topic that are reviewing [18].

According to Peer Review in Environmental Technology Development Program by Committee on the Department of Energy-Office of Science and Technology [18], there are two characteristic that measure the independency of the peers, and they are as follows,

- The individual or group do not involved as a participant, supervisor, technical reviewer, or advisor to the work being reviewed.
- They have the freedom from consideration, in order to make sure the topic is equally reviewed.

For this study, peer review is done by giving questionnaire to the students who already hold a Diploma in Computer Science and Mathematics, which also students of Degree in Computer Science and Mathematics, Hons Multimedia Computing from the Faculty of Computer Science and Mathematics, UiTM Jasin, Malacca, Malaysia. The purpose of choosing them is because the need for people who are very familiar with the multimedia surroundings. Their computing and multimedia background helps a lot in justifying the review.

Estimation for the sample size of peer review is a crucial need for an effective and reliable result. Calculating margin error to measure the confidence interval of the questionnaires can be done using the formula in Fig. 5 [19].

$1 / \sqrt{N}$
Where <i>N</i> is the number of participants or sample size

Fig. 5. Margin error formula

For example, if there is a 95% of confidence interval, meaning there is only 5% of margin error. The further details of how much sample size is needed parallel to the margin error is already calculated in the Table I, below.

Table I. Margin error for each sample size

Sample Size (N)	Margin Error	Margin Error (%)
10	0.316	31.6
20	0.224	22.4
50	0.141	14.1
100	0.100	10.0
200	0.071	7.1
500	0.045	4.5

1000	0.032	3.2
2000	0.022	2.2
5000	0.014	1.4
10000	0.010	1.0

Based on the Table I, it is observable how much margin error will occur for every different sample size. For a sample size of 10, the margin error is quite high with 31.6%, compare to sample size of 10,000 with margin error of only 1%. It is observable that the higher number of sample size contribute to a lesser percentage of margin error, and higher confidence interval. For this study, peer reviews with sample size of 20 are used, with confidence interval of almost 80%.

Data is gathered and analyzed from the questionnaire. A questionnaire set from Chertoff (2010) is used as a reference and only the variables are modified to meet with the functionality of FdLM 3D Walkthrough.

There are four categories in the questionnaire. Table II shows division of questionnaire is to ensure a more specific and focus questionnaire session.

Table II. The category division of questionnaire

Categories	No of Questions
Interface	4
Virtual reality elements	4
Interactivity	1
Conclusion	1

The questionnaire was conducted by asking the respondents to go to the official Malacca Maritime Museum virtual tour at <http://www.virtualmuseummelaka.com/maritime.htm>. Respondents were then asked to have a look through of the existing virtual tour.

It took about ten to twenty seconds to get attention from the users when they are browsing a website but an average of two minutes is enough for a user to interact with the website [20]. In the questionnaire session, respondents were given ten minutes to interact with the FdLM virtual tour. A longer minutes is purposely given so that the respondents able to carefully examine the existing virtual FdLM.

The result of the questionnaire is documented and analyzed as in Table III. The mean of respondents rating from each question is gathered and analyzed.

Table III. Mean of Interface questions

No	Interface Questions	Mean
1	The virtual tour reflects a detail interior design of the FdLM Ship	1.75
2	The graphic used to represent FdLM is in high definition (HD)	1.35

3	The slide shows give a good graphic representation of FdLM	1.35
4	The overall interfaces meet my expectation of how a virtual reality application should look like.	1.35

The questionnaire starts with question one asking if the interface of the virtual ship enable the respondents to understand the interior design of FdLM. This is to ensure if the interface provided gives significant value to the detail of FdLM. The frequency of response results in eight respondents rate one as strongly disagree, nine respondents rate two as disagree, while the other three as moderate. The rating resulted in mean of 1.75.

The second question emphasis on the viewing pleasure of the virtual ship, whether the respondents able to experience high definition graphic. This multimedia element is vital in ensuring a clear image while browsing through the virtual ship. Question two has mean score of 1.35 where thirteen respondents rate strongly disagree and seven respondents as disagree for the high definition aspect of the graphic used in the existing virtual ship. This shows that the existing FdLM virtual tour needs improvement in term of HD graphic.

Meanwhile, question number three would ask respondent to rate the slide show provided in the existing virtual ship, if it is able to give a good graphic representation for the ship. The main purpose for this question is to know whether the slide show act as a good interface medium in describing the FdLM. However, twelve respondents rate strongly disagrees, while the other eight disagree with the statement. Thus, result in mean score of 1.35. The mean score shows that the slides show does not help in representation of FdLM.

As all of the respondents have basic in multimedia studies, it is important for them to rate the existing virtual ship whether it meet their expectation of a virtual reality tour. Based on the questionnaire result for question four, it is observable that thirteen respondents rate strongly disagree, while the other seven rates disagree that the existing virtual ship interface meet their expectation. The rate given resulted with mean of 1.35. As a conclusion, the respondents strongly disagree that the overall interface of existing FdLM tour meet their expectation of a VR application.

Table IV. Mean of VR questions

No	VR Questions	Mean
5	I am able to feel the presence of object inside the FdLM from the virtual museum.	1.25
6	The virtual museum has a high quality of immersion.	1.20
7	I am able to imagine how FdLM looks like in real life.	1.30

8	The overall virtual reality elements meet my expectation of how a virtual application should look like.	1.35
---	---	------

The second section of the questionnaire is regarding virtual reality elements as in Table IV. The fifth question would like the respondents to rate the realness of the existing virtual ship as the main purpose of virtual reality is to let the user believe that they are actually in the environment they are experiencing. Fifteen respondents vote strongly disagree and the other five votes disagree. Question five has the mean of 1.25. This indicate that the majority of respondents not able to feel the presence of object inside the FdLM from the virtual museum.

Question six has the mean of 1.20 with frequency of response sixteen respondents vote strongly disagree and the other four as disagree. While for question seven, it is scored with the mean of 1.30, with fourteen respondents vote as strongly disagree and six as disagree.

The last question for virtual reality, focus on the respondents' opinion of the existing FdLM virtual reality meet their expectation of how a virtual application should be as show in Fig. 6. Question eight score the mean of 1.35 with thirteen respondents' rate as strongly disagree and seven respondents disagree.

Table V. Mean of Interactivity question

No	Interactivity Question	Mean
9	I am able to interact well with the multimedia elements from the virtual museum. (ex: click through links and photos etc)	1

The main purpose of question number nine is to identify if respondent able to interact well with the existing virtual ship as in Table V. The current virtual ship, would not even allow the respondent to click between photos or zoom in from the photos in the slide show provided. The virtual application only allows respondents to view the photos while reading some text. The photos were automatically looped and it is display from the slide show. Thus, it results in the mean of 1 where all respondents rate this question as strongly disagree for them to interact with the multimedia elements from the virtual museum.

Table VI. Mean for Conclusion question

No	Conclusion	Mean
10	The existing virtual museum is in need of improvement especially in term of multimedia elements.	4.6

Question ten in Table VI emphasis more on the respondents' opinion whether they think the existing virtual FdLM need any improvement or otherwise. Based on the Mean, majority of respondents agree that the current FdLM virtual museum need improvement especially in term of multimedia elements.

As a conclusion, the existing virtual museum for FdLM does need a lot of improvement. A more user friendly and interactive user interface should be implemented for a more entertaining experience. In addition, modelling 3D object for FdLM would help user to feel more immerse in the virtual environment. At the same time a 3D Walkthrough will act as a visual aid for them to imagine and understand FdLM in a deeper perspective.

C. Design and Development

Navigational map provide specific information with links that show how those information are connected and interact with each other. Fig. 6 below shows the navigational map for FdLM 3D Walkthrough.

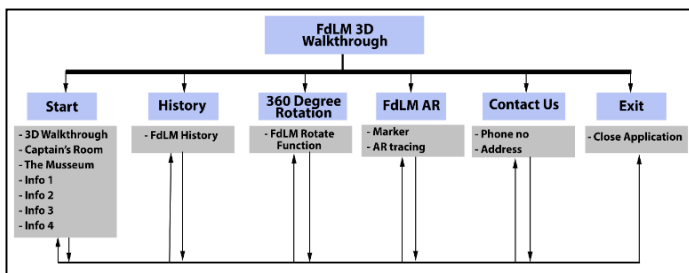


Fig. 6. Navigation map

There are six scenes from the FdLM 3D Walkthrough. The 3D Walkthrough will be at the Start scene. The History scene will introduce user to brief historical information of the FdLM ship. In order to help the user understand the ship better, they are able to see the ship in 360 degree rotation. This will help the user in knowing specifically the design of the ship. Next, an element of AR is implemented at the FdLM AR scene to let the user imagine the ships' design. This is done by connecting the application with the camera. Contact information of the Maritime Museum is available at the Contact Us scene.

Interface is important in developing an application. Interface in general should be in prospect of when a user give command to the computer, it responds back in a manner of showing what the user needs. User interface is much related to user interaction which can be delivered through input devices such as keyboard, mouse, touch screen and microphone [21]. Interface also influence in a good user experience and act as a major ingredients for a successful application today [22]. Thus, it is advised to use keyboard and mouse as a primary input devices for FdLM 3D Walkthrough since it is a desktop application.

Intel Developers has come up with desktop user interface guidelines along with the launch of Windows 8 in 2012 for a better desktop experience for their customers. Based on those

guidelines, FdLM 3D Walkthrough is designed to fit into desktop application as shown in Fig. 7.



Fig. 7.Division of Desktop Interface

There are three separated partitions with one blue partition and two pink partitions. The blue partition represent the main focus or the main content of a desktop application while the other two pink partitions are made for interaction purpose such as icons and buttons.

Besides the desktop user interface; colors also play a vital role in catching the attention from user. A complementary color method is used in the application, especially when it comes to icons. Fig. 8 shows the Complementary Color Chart. For example, a background with brown color that situated between the yellow and orange will be complement with an icon that has blue color.

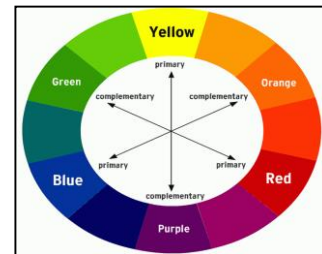


Fig 8. Complementary Color Chart

The complementary chart resembles how a developer should play with color in their application. A high contra of colors, for example orange with blue, would create a vibrant interface especially in full saturation. Complementary color also helps user to focus to a specific icons [23].

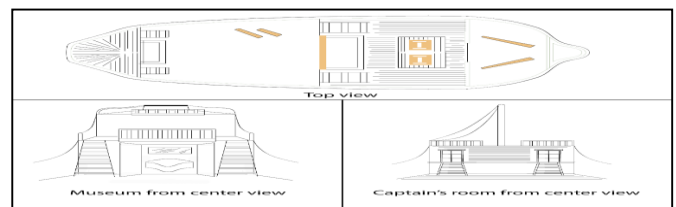


Fig. 9. Sketch of FdLM Ship

Sketching of the FdLM ship helps in specifying its look in the application. It would also give a brief view of the interior design of the ship. Development of the ship 3D model is referred to the sketch.

There are three views for the FdLM sketch as shown in Fig. 9. The top view shows the design of the ship from above, and two center views that show the Museum and Captain's Room.

The development of 3D modelling is done using 3D Studio Max. 3D Studio Max is chosen based on its capabilities and ease of use to develop 3D model and its compatibility to transfer file to other software. Fig. 10 shows the modelling process of FdLM.

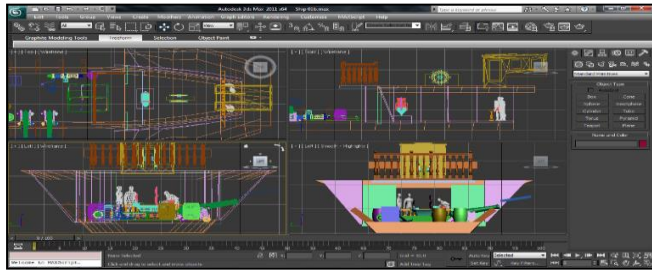


Fig. 10 Modelling Process

As modelling 3D object is the basic step of developing this application, several techniques is used in order to design the iconic ship. Some model only required to use one technique, while many other 3D objects required combination of two or more techniques. Listed in Table VII are some of the techniques used.

Table VII. Techniques used in modelling FdLM ship.

No	Techniques
1	Line tool
2	Extrude
3	Intrude
4	Slice plane
5	Editable poly/mesh with vertex, border, and polygon.
6	Texturing

Next, the finished 3D model is then converted to fbx file to be transferred to 3D Unity. Interaction and modelling setup is done within the virtual reality environment. Designing and development of the virtual surrounding happens in 3D Unity. Repositioning the 3D model is needed to place it at the right place. Incase if the model that we imported does not match with the VR environment like how it should be, adjusting the position, rotation and scale can be done at the Transform panel. Additional multimedia elements such as links and audio are inserted to enhance the look and feel of the ship.

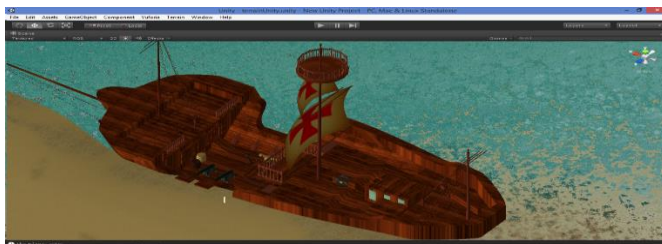


Fig. 11. FdLM from bird's eye view



Fig. 12. FdLM Information Screen

Fig. 11 and 12 shows the application interface that has been developed and enhancement are made based on the data gathered from initial investigation.

D. Implementation of Flor de La Mar (FdLM)

Implementation is done to know what type of hardware and software are needed in order to develop the project parallel with its functionality. Implementation process is where the design plan will be delivered into multimedia program. The suitable software and hardware selected and used in development of FdLM 3D Walkthrough.

E. Testing

The same 20 respondents from the previous peer review are called back to evaluate the FdLM 3D Walkthrough. The purpose of this testing is to know if FdLM 3D Walkthrough has offered improvement in term of multimedia elements compare to the existing virtual museum. Table VIII shows the mean result for interface questions.

Table VIII The mean result of FdLM Interface

No	Interface Question	Mean
Q1	The 3D Walkthrough reflects a better interior design of the FdLM Ship compare to the existing virtual museum.	4.00
Q2	The graphic used to represent FdLM is in high definition (HD)	4.10
Q3	The 3D models give a good graphic representation of FdLM	3.95
Q4	The overall interfaces meet my expectation of how a virtual reality application should look like.	4.40

The total mean score of this evaluation questionnaire is 5. For question one, its purpose is to know if the 3D Walkthrough gives significant value to the detail of FdLM ship. From the Table VIII, we could see the rating for Q1 is 4.00 out of 5. This means that the respondents do agree that the 3D Walkthrough reflects a better interior design of the FdLM Ship compare to the existing virtual museum.

As for Q2, the mean score is 4.10 out of 5. This shows that there are improvement for the representation of FdLM in term of high definition when using 3D Walkthrough. For Q3,

this question would like to evaluate if the 3D model use is effective in representing FdLM, and it scores the mean of 3.95. Meanwhile, the overall interface scores 4.40 out of 5, which means the respondents do agree that the interface used meet their expectation of how a virtual reality application should look like.

Table IX The testing result for VR

No	Virtual Reality Questions	Mean
Q5	I am able to feel the presence of object inside the FdLM from the 3D Walkthrough.	4.50
Q6	The 3D Walkthrough has a high quality of immersion.	4.55
Q7	I am able to imagine how FdLM looks like in real life.	4.70
Q8	The overall virtual reality elements meet my expectation of how a virtual application should look like.	4.60

Next, Q5 evaluates if the user able to feel the presence of object inside the FdLM and it is proven so when this question score 4.50 out of 5. For the immersion element of VR, question 6 score 4.55 which really stated that the 3D Walkthrough does effect the user in term of immersion. For Q7, it scores 4.70 out of 5 and this shows that majority of the user able to imagine how FdLM looks like in real life. As for the last question in the VR category, Q8 emphasize on the overall VR elements from the 3D Walkthrough and it has the mean of 4.60. This means that majority of the respondents do agree that the FdLM 3D Walkthrough meet their expectation of how a virtual application should look like as shown in Table IX.

Table X shows the testing result of Interactivity

No	Interactivity	Mean
Q9	I am able to interact well with the multimedia elements from the FdLM 3D Walkthrough. (ex: click through links and photos etc)	4.30

As shown in Table X for interactivity elements, Q9 evaluate of how the interactivity of the 3D Walkthrough plays its role. Interactivity from the application able to score 4.30 out of the total mean score of 5. This shows that the majority of respondents do agree that the FdLM 3D Walkthrough offer a good interactivity with multimedia elements.

Table XI shows the mean of Conclusion question

No	Conclusion	Mean
Q10	I prefer FdLM 3D Walkthrough compare to the existing virtual museum.	4.65

For the overall aspect of 3D Walkthrough, a conclusion question is asked to know the opinion of the respondents. The majority of respondents do agree that to prefer FdLM 3D Walkthrough compares to the existing virtual museum with mean rate 4.65 out of 5 as shown in Table XI.

IV. CONCLUSION

Virtual reality has a big opportunity to be explored especially in the tourism industry. Developing a VR application for FdLM is just a stepping stone to promote Malacca and its historical sites. Ana Serrano, a Chief Digital Officer of the Canadian Film Centre (CFC), and founder of CFC Media Lab, in her talk during TED events define virtual reality as a rich, visual, multi-sensory computer simulated environment in which user could emerge themselves in and also interact. Based on the testing result, it is observable that FdLM 3D Walkthrough serves a better interface, interaction and virtual reality element compare to the existing FdLM virtual tour.

As for the future work, the application will be tested to the random visitors of FdLM museum. The purpose of this is to get a public insight of how the application able to help them in guiding and promote understanding of the historical ship.

In conclusion, FdLM 3D Walkthrough has greatly facilitate historical understanding using multimedia elements. The application also helps user to be more immersive, imaginative and interactive, thus, offer solution to the existing FdLM virtual museum.

REFERENCES

- [1] S. S. A. H. Mustapa, *Showcasing Maritime Heritage Artefacts for the Benefit of the Tourist Industry in Malaysia*. International Journal of Nautical Archaeology 34(2): 211-215, 2005.
- [2] D. Ince, "Acoustic coupler," in A Dictionary of the Internet. Oxford University Press, [online document], 2013. Available: Oxford Reference Online, <http://www.oxfordreference.com> [Accessed: Jan 24, 2015].
- [3] J.Vince, *Introduction to Virtual Reality*. Springer Science & Business Media, 2014.
- [4] S. A. Aukstakalnis, et al, *Silicon Mirage: The Art and Science of Virtual Reality*. Peachpit Press, 1992.
- [5] G. Burdea, & P. Coiffet. *Virtual reality technology*. Hoboken, NJ: J. Wiley-Interscience, 2003.
- [6] M. Z. Rani, *Assessing customer behavior towards tourism website in Malaysia*, MARA University of Technology, 2009
- [7] S.L. Hsu, & J. C. C. Lin, *Acceptance of blog usage: The roles of technology acceptance, social influence and knowledge sharing motivation*. Information & management, 45(1), 65-74, 2008.
- [8] P. Raman, *For a virtual tour of Malaysia*. New Straits Times Press (Malaysia), 2002.
- [9] C.Y. Chan, *Virtually virtual*. New Straits Times Malaysia, 2000.
- [10] R. Azuma, *A survey of augmented reality*. Hughes Research Laboratories, 1997.
- [11] R. Azuma. et al, "Recent advances in augmented reality," *Computer Graphics and Applications, IEEE*, vol.21, no.6, pp.34-47, 2001.

- [12] M.E.C. Santos, et al *Augmented Reality Learning Experiences: Survey of Prototype Design and Evaluation*, Learning Technologies, IEEE Transactions on , vol.7, no.1, pp.38-56, 2014.
- [13] P. Milgram and F. Kishino. *A Taxonomy of Mixed Reality Visual Displays*, IEICE Trans. Information and Systems, cvol. 77, no. 12, pp. 1321-1329, 1994.
- [14] B. Furht, *Handbook of Augmented Reality*, Springer, 2011.
- [15] M. Hirzer, *Marker detection for augmented reality applications*, Inst. for Computer Graphics and Vision Graz University of Technology, Austria, 2008.
- [16] C. Owen, et al *Comparative Effectiveness of Augmented Reality in Object Assembly*. Proceedings of the ACM CHI 2003 Human Factors in Computing System (CHI2003) pp. 73-80, 2003.
- [17] WordNet 3.0, Farlex clipart collection. S.v. "site visit." Retrieved February 2 2015 from <http://www.thefreedictionary.com/site+visit>
- [18] C. D. E. O. S. T. P. R. Program, and N. R. Council *Peer Review in Environmental Technology Development Programs*, National Academies Press, 1999.
- [19] R. Niles, *Robert Niles' Journalism Help: Statistics Every Writer Should Know*, RobertNiles.com. Retrieved June 31, 2014 from <http://www.robertniles.com/stats/>, 2006.
- [20] J. Nielsen, *How long do users stay on web pages?*. Retrieved November 28, 2014 from <http://www.nngroup.com/articles/how-long-do-users-stay-on-web-pages/>, 2011.
- [21] R. Lai, *Digital Design Essentials: 100 ways to design better desktop, web, and mobile interfaces*. Rockport Publishers, 2013.
- [22] M. Rao, *User Interface Design Guidelines for Great Experience Design*. Intel Developers, 2012.
- [23] S.J. Wenrich, *All the colors of life : From the mystery and history of color and secret*

Enhanced Watermarking Scheme for 3D Mesh Models

Lamiaa Basyoni

College of Computing and Information Technology
Arab Academy of Science and Technology
Cairo, Egypt
lamiaa.basyoni@gmail.com

H. I. Saleh

Radiation Engineering Department
Atomic Energy Authority
Cairo, Egypt
h_i_saleh@hotmail.com

M. B. Abdelhalim

College of Computing and Information Technology
Arab Academy of Science and Technology
Cairo, Egypt
mbakr@ieee.org

Abstract— In this paper we present a new non-blind watermarking scheme for 3D graphical objects (meshes). Non-blind watermarking scheme is known to be more secure than blind ones, since the original and watermarked models are needed for extraction. In our scheme we use the model's prominent feature points to divide the model into separate segments, these feature segments are then projected from 3D representation to the 3 main 2D-Planes. The watermark embedding is done in frequency domain of these projections. The experimental results showed the robustness of this scheme against various mesh attacks (mesh simplification, subdivision, smoothing, cropping, etc). This scheme also allows quite large payload to embed. The results of the proposed scheme showed average of 50 percent improvement in robustness against geometry attacks, and up to 70 percent against connectivity attacks.

Keywords— *Digital Watermarking, 3D Models, Triangular Mesh, Robust Watermarking.*

I. INTRODUCTION

3D mesh models are rapidly growing in the multimedia applications, and are used more and more in many fields, with industrial, medical, and entertainment applications. This increases the need for intellectual property protection and authentication.

Digital watermarking was found to be a very efficient solution for the authentication problems. This technique carefully hides some secret information in the functional part of the cover content. A watermark is a digital code permanently embedded into a cover content [1]. A watermark can be embedded in a variety of cover content types, including images, audio data, video data, and 3D graphical objects.

The 3D graphical objects are the most difficult kind of digital media to design a watermarking framework for, as it has many challenges [2], such as: (1) Low volume of data: the

amount of data available to hide the watermark in it is very low as a 3D model consists of a few thousands of vertices unlike the enormous amount of pixels provided in the case of images. (2) No unique representation: an image is represented as a 2D array, while a 3D model can be represented in many different ways. (3) No robust transformation field that can be used for embedding. (4) Attacks may change the geometry and connectivity properties of the mesh. (5) High computational requirements, specially for frequency domain implementation.

Watermarking techniques are generally classified based on the detection method to blind and non-blind techniques. Blind techniques require neither the cover (original model) nor the embedded watermark to extract the watermark, while the non-blind techniques need the cover content in order to complete the extraction process, so that, the possession of the original model becomes part of the proof of ownership[3]. The non-blind watermarking is more robust than the blind watermarking. Both techniques gained a lot of attention

recently, many algorithms were developed to provide robust watermarking scheme [4].

In the first category, a recent blind watermarking technique was developed based on multi resolution representation and fuzzy logic. Fuzzy logic approach approximates the best possible gain with an accurate scaling factor so that the watermark remains invisible. The fuzzy input variables are computed for each wavelet coefficient in the 3D model. The output of the fuzzy system is a single value which is a perceptual value for each corresponding wavelet coefficient. Thus, the fuzzy perceptual mask combines all these non-linear variables to build a simple, easy to use HVS (human visual systems) model. Results showed that the system is robust against affine transformations, smoothing, cropping, and noise attacks [5].

Another blind watermarking scheme based on volume moments was introduced by Wang, et al [6]. During watermark embedding, the input mesh is first normalized to a canonical and robust spatial pose by using its global volume moments. Then, the normalized mesh is decomposed into patches and the watermark is embedded through a modified scalar Costa quantization of the zero-order volume moments of some selected candidate patches [6].

In the second category, there is the non-blind watermarking algorithm based on geometrical properties of 3-D polygon mesh introduced by Garg, H., et al [4]. The objective of this algorithm is to process the object to find the less visible area of 3D polygonal mesh. Another non-blind watermarking scheme that was developed by Ryutarou Ohbuchi, et al [7], uses the spectral domain to embed the watermark. The algorithm computes spectra of the mesh by using eigenvalue decomposition of a Laplacian matrix derived only from connectivity of the mesh.

The rest of the paper is structured as follows: section 2 provides an overview of the segmentation operations we apply on the mesh model. Section 3 describes the steps of embedding and extracting the watermark. In Section 4 we illustrate the main features of our implementation. The experimental results are provided in section 5, and a conclusion of the paper is presented in section 6.

II. SEGMENTATION

The first step in our algorithm is to divide the original mesh model into core part, and a number of segments contain its feature points. In this section we describe the used segmentation method based on [8]. The feature points selected



Fig. 1. Prominent Feature Points

are the prominent points of the model. These points reside on the tip of prominent components of the model. For instance, in Fig.1, feature points can be found on the tip of the tongue, horn, and tail. To formally define the vertices on the tips, it should satisfy the following conditions.

$\forall v \in S$, let N_v be the set of neighbouring vertices of vertex v . Let $GeodDist(v_i, v_j)$ be the geodesic distance between vertices v_i and v_j of mesh S . The local condition that a feature point should satisfy is that $\forall v_n \in N_v$.

$$\sum_{v_i \in S} GeodDist(v, v_i) > \sum_{v_n \in N_v} GeodDist(v_n, v_i) \quad (1)$$

The feature points determined are then used to guide the segmentation. The mesh is segmented into its core component and its prominent components. Each prominent component is defined by one or more of the feature points, while the core components are closer to the center of the mesh model. The segmentation process consists of the following 3 steps:

1) Spherical Mirroring:

Prominent feature points on surface S tend to be extreme in some direction, while vertices of the core component tend to be closer to the center of S , the aim of spherical mirroring is to reverse this situation, and the vertices of the core become external and can be easily extracted.

2) Core Component Extraction:

The convex hull of the mirrored vertices is computed. The vertices that reside on the convex hull, along with the faces they define on S , are considered the initial core component.

3) Extraction of the other segments:

Once the core component is found, the other segments of the mesh are extracted by “subtracting” the core component from the mesh.

A. Segments Projection

The term *projection* refers to any dimension-reduction operation. One way to achieve this is using a scale factor of zero in a certain direction, thus, all points will be projected onto the perpendicular plane (in our 3D case). This type of projection is called Orthographical Projection or Parallel Projection as shown in Fig.2. In our framework, we use parallel projection to convert the feature segments selected of the 3D mesh model into three 2D-arrays, each obtained by projecting against one of the 3 cardinal axes.

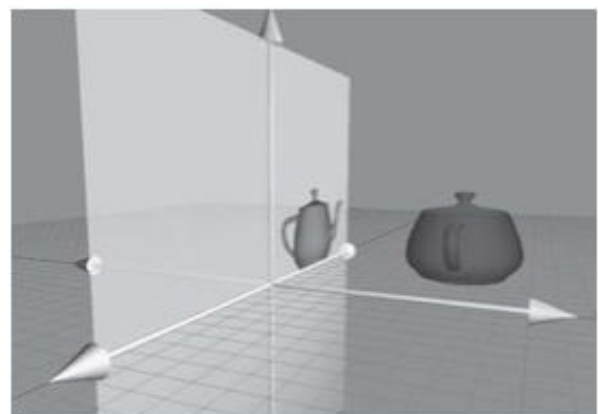


Fig. 2. parallel Projection of a 3D object

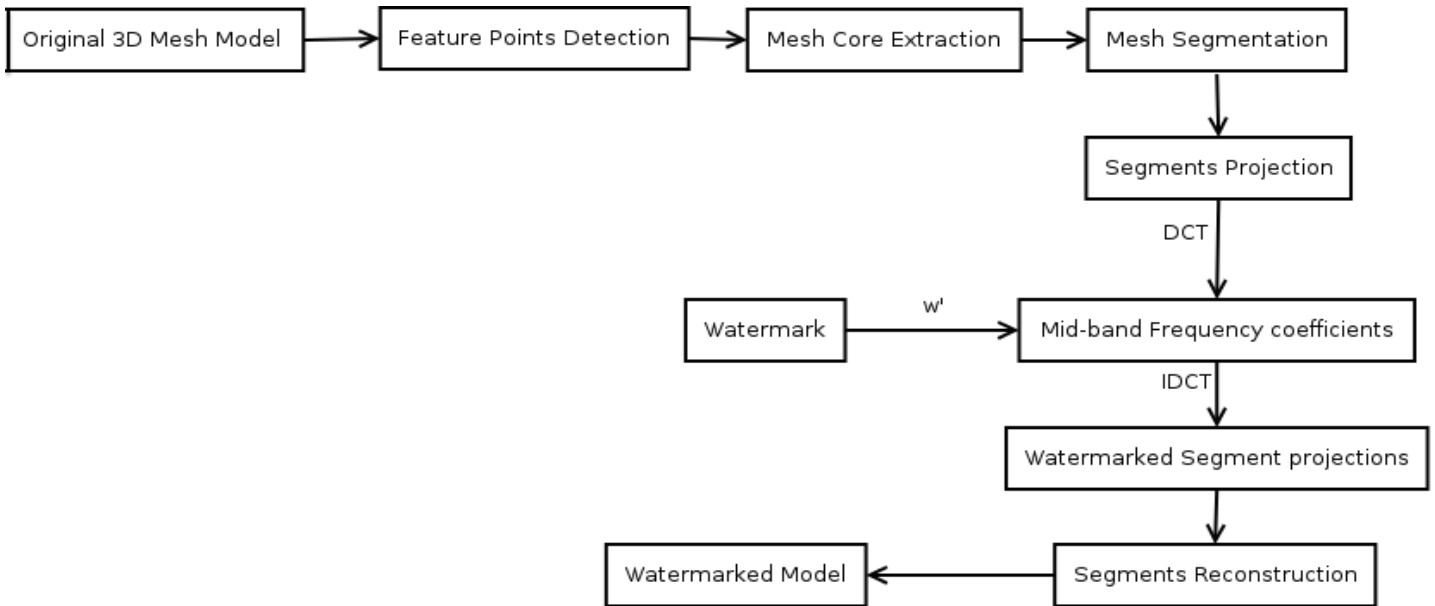


Fig.3. Steps of the watermarking embedding process

To project a 3D model or a segment of it onto a plane, we use a scale value of zero on the perpendicular axis to this plane [8]. The 3D matrices used for projection on the xy, xz, and yz planes are as follows:

$$P_{xy} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$P_{xz} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$P_{yz} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

III. WATERMARK EMBEDDING AND DETECTION

A. Watermark Embedding Process

The watermark embedding process is based on the segmentation process, and the segments projection explained in the previous section. Fig.3. shows the scheme we use for watermark embedding process. In our scheme we use a sample image as watermark payload. The pixels of the image (with values 0 and 1 only) are embedded in the selected segments defined by the feature points of the model. The watermark image is split into equal blocks of pixels, as shown in Fig.4. Each block is then embedded in a segment of the 3D model.

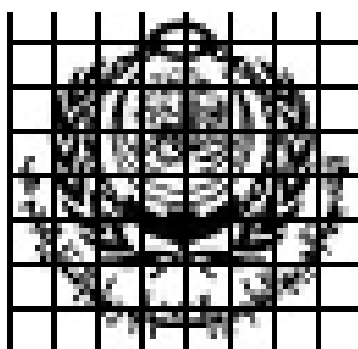


Fig. 4. Watermark Image split into blocks

Dividing the image watermark into blocks and distribute it among the feature segments increase the resistance for attacks like cropping. It has been proved in cases of image, audio, and video watermarking that it is better to embed information in spectral domain rather than in the spatial domain based on [1]. Many of the new researches focus on the use of frequency domain for its robustness compared to other domains [14]. Robustness is a key factor in our algorithm, so we're adopting the DCT (discrete cosine transform)-based watermarking method. DCT is selected for its computational simplicity compared to other transforms such as Discrete Fourier Transform. It also has the ability to pack more information in fewer coefficients. The DCT coefficients are divided into 3 main bands; low frequencies, mid frequencies, and high frequencies, as shown in Figure 5. Embedding in the mid-band coefficients avoid scattering the watermark information to most visual parts of the model i.e. the low frequencies and also it do not overexpose them to removal through noise attacks where high frequency components are targeted.

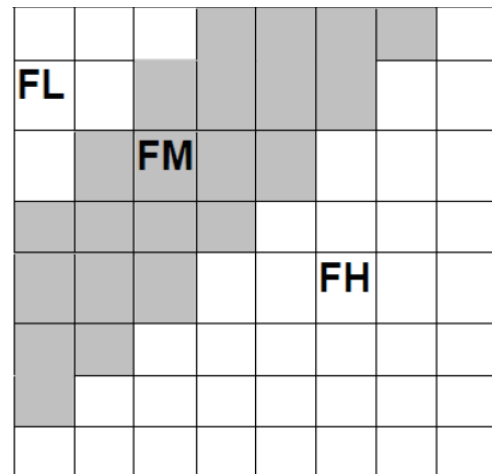


Fig. 5. DCT coefficients for 8*8 block

The watermarking process can be described as follows:

- The arrays resulting from the projection of the feature segments are divided into blocks, and DCT is applied on each block
- A number (N) of the mid-band coefficients is selected, where $N = 1/4$ the width of the watermark image.
- The embedding is then performed in the transform domain; the logo image we use is represented as 0's and 1's. If the pixel value to be embedded is 0, then there will be no changes in the coefficient value: $C' = C$. Otherwise the

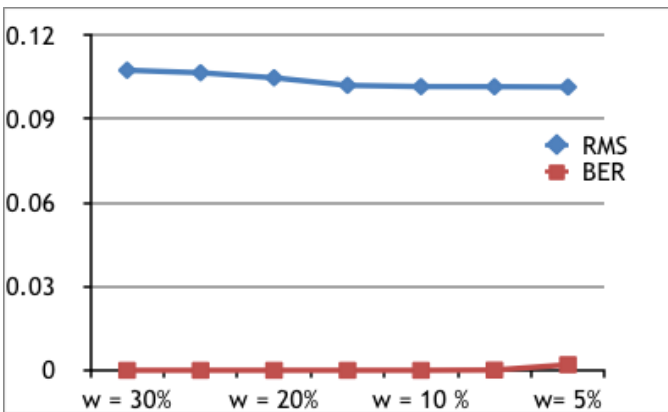


Fig. 6. The effect of different watermark ratios on visual quality and the Bit Error Rate

coefficient will be changed to $C' = C + w \cdot C$, where C' is the DCT coefficient after embedding, and w is the watermark ratio, that has a direct effect on the visual quality of the model. Figure 6 shows the effect of the watermark ratio w on the visual quality measured by the Root Mean Square error (RMS) and the bit error rate of the watermark after extraction. At 7% of the coefficient values we reach an acceptable RMS value (according the watermarking benchmark [13]), and the watermark can be fully recovered.

- IDCT (inverse discrete cosine transforms) is then applied to these blocks to generate the watermarked segments projections.
- We then reverse the projection steps to re-create the watermarked segments.
- 3D object reconstruction by combining the core part with the watermarked feature segments.

B. Watermark Extraction Process

Fig.7. shows the block diagram of watermark extraction process. The process has the same steps to create the segments projection, and since we're adopting a non-blind technique, the

original segments projections are needed to complete the extraction process. We use the same process, whether the watermarked model is attacked or not.

The array of segments projection generated for both watermarked and original models will be the input for DCT operation. The same range of mid-band coefficients is selected to extract the watermark:

$$W_L = \frac{C_W - C_O}{\beta \cdot C_O} \tag{2}$$

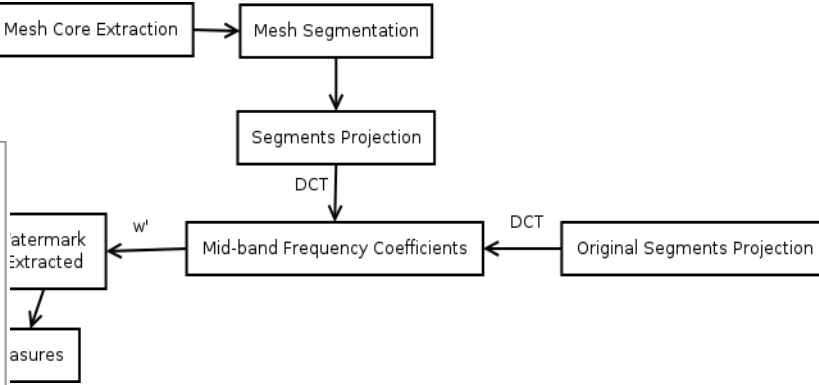


Fig. 7. The watermarking extracting process

Where C_w is the coefficient obtained for the watermarked model, C_o is the original model coefficient, and W_L is the detected value for the corresponding pixel in the logo image, it should have the value of 0 or 1. This way the logo image will be re-constructed. The original image will be compared to the extracted image using the PSNR (peak signal to noise ratio) measure to calculate the efficiency of our watermarking scheme.

IV. EXPERIMENTAL RESULTS

A. Distortion Evaluation

The described scheme was applied on different 3D models whose characteristics are listed in Table 1.

Table.1. Characteristics of the 3D models used in experiments.		
Object	No. Vertices	No. Faces
Bunny	34835	69666
Dragon	50000	100000
Hand	36619	72958
Rabbit	70658	141312
Venus	100759	201514

The results of applying our scheme on the five objects are shown in Fig.8.

The models were selected to provide a diversity of mesh shapes; a shape like the bunny has many rounded faces, where

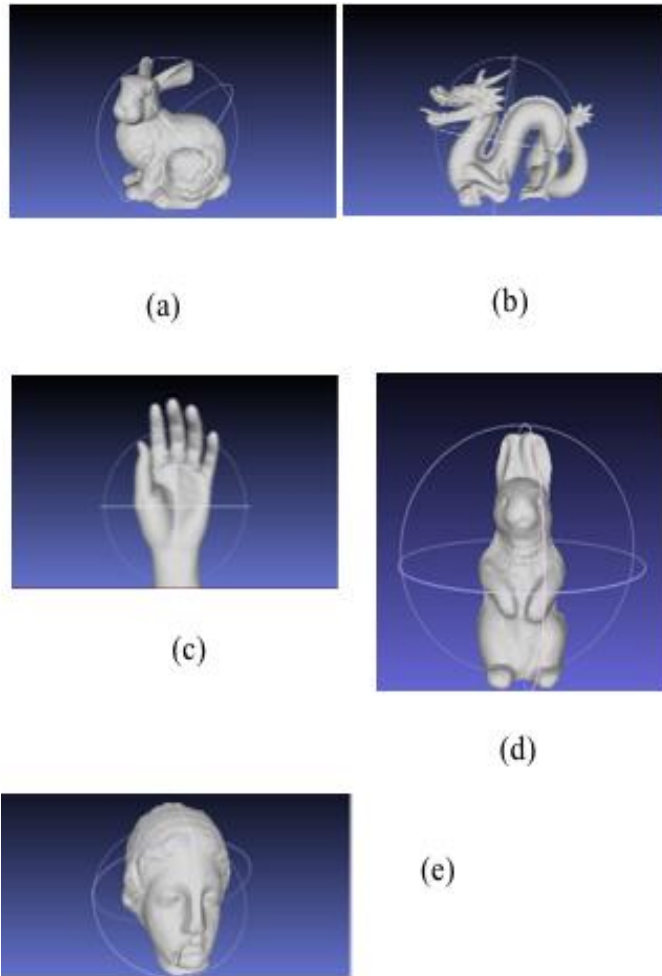


Fig. 8. 3D models used in our experiments: (a) bunny, (b) dragon, (c) hand, (d) rabbit, and (e) venus.

the dragon shape is very complex. There is also an elongated object (the rabbit), and the hand object is quite flat.

The watermark embedding process introduces some distortion to the original cover mesh. This distortion can be measured using many metrics; we are using the root mean square error (RMS) between two 3D-surfaces which is more accurate than other simple vertex-to-vertex distance measures (e.g. PSNR). The RMS is defined in eq.3.

$$d_{RMS}(S, S') = \sqrt{\frac{1}{|S|} \sum_{p \in S} d(p, S')^2} \quad (3)$$

Where p is a point on surface S , S' is the surface to measure the distance to, $|S|$ is the area of S , and $d(p, S')$ is the distance between p and S' . The amount of distortion introduced by a watermarking technique is one of the evaluation factors; it's required to present the minimal amount of distortion. The method used to measure the quality of the watermarked models is Metro [10]. Mesh models are usually used in digital entertainment applications, so it has to be assured that the embedding of the watermark will not affect the visual quality of the models. Based on the evaluation criteria defined by the benchmark of 3D models watermarking [11] the induced geometric distortion should be <0.09 with respect to the diagonal of the bounding box. Table.2 shows the baseline evaluation results of the proposed scheme and Wang's algorithm [12], stating the payload size used with every model and the visual effect of it in terms of the perceptual quality measure RMS. The evaluation shows that our scheme is supporting much larger payload, while maintaining the models visual quality. In Table.3 the ratio between payload and RMS is presented to show that our scheme is better by 36% in preserving the visual quality of the 3D Model for large watermarking payloads.

Model	Payload		RMS(w.r.t. lbbd)	
	Wang's	Proposed Scheme	Wang's	Proposed Scheme
Venus	75	256	0.0023	0.0027
Bunny	67	144	0.0017	0.0053
Horse	46	64	0.001	0.0014
Dragon	49	256	0.0018	0.0057
Average Values	59.25	180	0.0017	0.0038

	Wang's	Proposed Scheme
Payload/RMS	34852	47368

The visual effects of embedding the watermark in the mesh models are illustrated by zooming in details of the models (Fig.9).

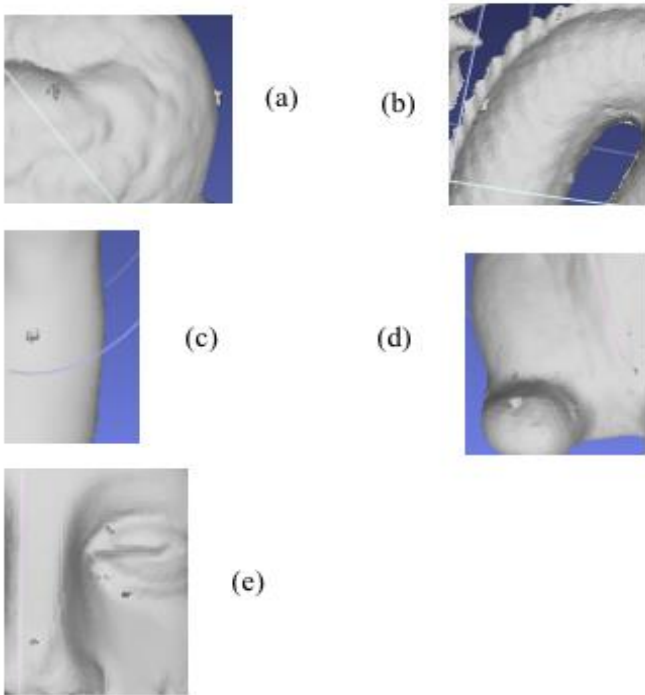


Fig. 9. Zoom-in area of the watermarked models shows the distortion added.

B. Information Embedding Capacity

In our proposed scheme we provide embedding capacity (watermark payload) that varies according to the model size and number of feature point (a logo image of 16x16 pixels, up to 24x24 pixels), which gives the ability to hide a considerably large data..

C. Robustness Evaluation

In this section we present the evaluation of our watermarking scheme against various 3D mesh attacks, in general there are three kind of routine attacks applied on watermarked meshes: *file attacks*, *geometry attacks*, and *connectivity attacks*. In the following we present the results of applying a diversity of these attacks, by measuring the amount of distortion introduced by these attacks on the watermarked model, and the quality of extracted watermark

- Cropping: one of the connectivity attacks in which one part of the watermarked mesh is cut off and lost, we applied this attack on the model where 12% of the mesh vertices were lost as shown in Fig.10 (a).

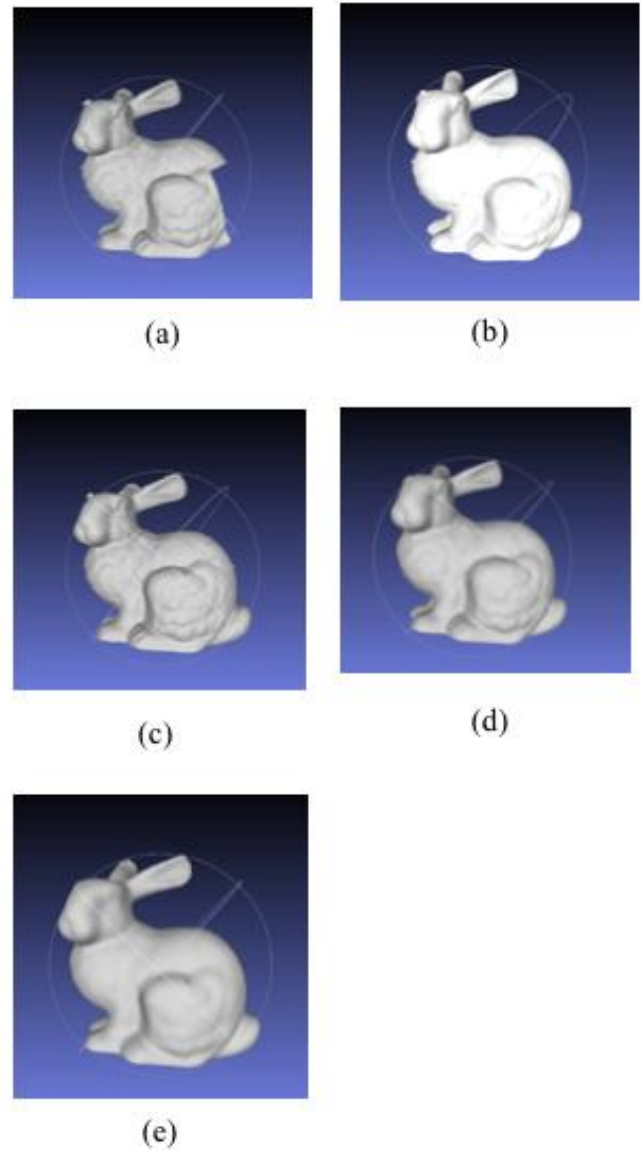


Fig. . The watermarked bunny model after applying 5 different attacks : (a) Cropping, (b) Subdivision using loop scheme, (c) Subdivision using mid-point scheme, (d) Laplacian Smoothing using 3 iteration, (e) Laplacian Smoothing using 10 iterations.

- Subdivision: a connectivity attack in which vertices and edges are added to the mesh to obtain a smoother and higher visual quality version of the model. Two different schemes of subdivision are used: the loop scheme [Fig.10 (b)], and the mid-point scheme [Fig.10. (c)].

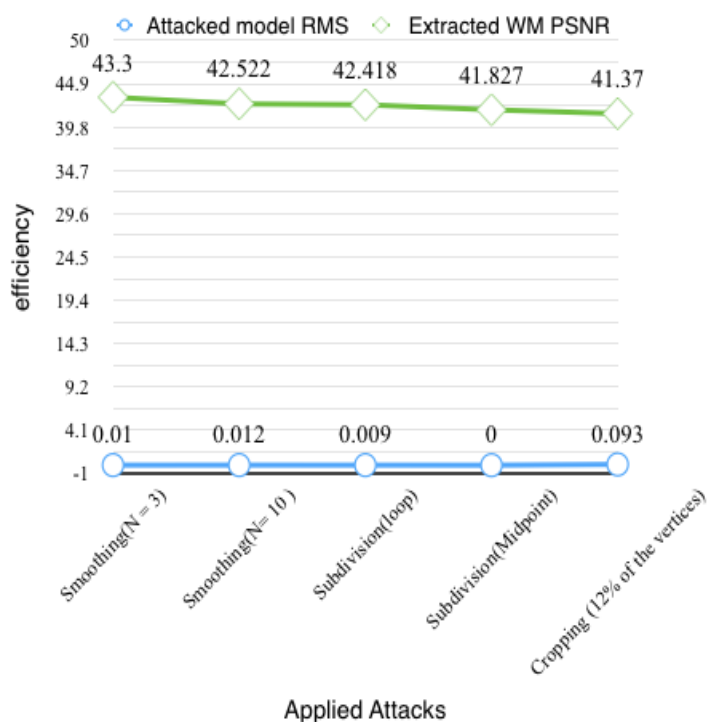


Fig. 11. The effects of some attacks on the visual quality of the watermarked model and the efficiency of watermark extraction.

- Smoothing: a geometry attack that's also a common process used to remove the noise generated during the mesh generation process. Referring to the benchmark we adopt [12] we applied Laplacian smoothing with different number of iterations [N = 3, N = 10], its effect is shown in Fig.10 (d) and (e) respectively. Applying these attacks on the watermarked mesh introduces different amount of distortion, and affects the hidden data in its own way. In Fig.11, we plot the effect of these attacks on the bunny model. The PSNR measures the quality of the watermark image restored after applying the attacks, the results show that the watermark is efficiently extracted from the attacked model.

V. RESULTS COMPARISON

In order to show the robustness of our scheme, it is compared to other schemes. Referring to the benchmark we used for evaluating the embedding effect on the models, it presents 2 different algorithms to compare with, the work of Wang et al. [11] that is based on modification of the mesh local volume moments, and the work of Cho et al. [12] that is based on modification of the mean value of the histogram of vertex norms.

Attacks	Chos's BER	Wang's BER	Proposed Scheme BER
Similarity Transformation	0.0	0.0	0.0
Smoothing N = 5	0.01	0.0	0.013
Smoothing N = 10	0.23	0.01	0.015
Smoothing N = 30	0.38	0.07	0.02
Smoothing N = 50	0.45	0.14	0.02
Average Geometry Attacks	0.214	0.044	0.014
Subdivision Midpoint	0.04	0.0	0.02
Subdivision	0.14	0.0	0.04
Subdivision Loop	0.16	0.0	0.045
Simplification E = 10	0.01	0.0	0.02
Simplification E = 30	0.05	0.0	0.019
Simplification E = 50	0.18	0.0	0.025
Simplification E = 70	0.33	0.0	0.029
Simplification E = 90	0.23	0.01	0.031
Cropping 10%	0.5	0.51	0.012
Cropping 30%	0.53	0.49	0.014
Cropping 50%	0.51	0.49	0.013
Average Connectivity Attacks	0.243	0.136	0.024

The benchmark limits the used payload to be around 70 bits in order to conduct a meaningful comparison, so we are reducing the size of our embedded watermark logo to an 8x8 image, which gives us 64 bits to be embedded. The perceptual protocol defined by the benchmark is applied, and the Bit Error Rate (BER) is calculated as a measure of robustness. Table 4 shows a comparison of the BER computed after applying a number of attacks on Venus model. Embedding the watermark in the middle frequency band does not expose it to removal by operation targeting the higher frequencies such as smoothing, and it can be noted that our scheme gives a remarkable robustness against this kind of attacks even at large number of iterations. The embedding of the watermark is also scattered between many feature segments that are not necessary interconnected, that makes it more robust against faces removing attacks, such as cropping.

CONCLUSION

In this paper, we proposed a new non-blind, frequency domain watermarking scheme that is based on mesh segmentation. The proposed scheme compared to other methods shows a better robustness against both geometry and connectivity attacks. The scheme also preserves the visual quality of the 3D mesh models. A considerably large payload is also supported, which allows hiding of large sum of information, and this one of the most critical issues in 3D models watermarking.

Watermarking with DCT Based Watermarking', *International Journal of Computer Theory and Engineering*, pp. 647–653, Jan. 2010.

REFERENCES

- [1] F. Yu, *Three-dimensional model analysis and processing*. 1st, ed. Heidelberg: Zhejiang University Press, 2007.
- [2] Y. Zhi-qiang, H. H. S. Ip, and L. F. Kowk, 'Robust watermarking of 3D polygonal models based on vertex scrambling', *Proceedings Computer Graphics International 2003*, pp. 254 - 257. 2003.
- [3] S. W. Foo, 'Non-blind audio-watermarking using compression-expansion of signals', *APCCAS 2008 - 2008 IEEE Asia Pacific Conference on Circuits and Systems*, Jan. 2008.
- [4] H. Garg, S. Agrawal, and G. Varshneya, 'A non-blind image based watermarking for 3-D polygonal mesh using its geometrical properties', *2013 Sixth International Conference on Contemporary Computing (IC3)*, pp. 313 - 318. 2013.
- [5] S. Tamane, 'Blind 3D Model Watermarking based on Multi-Resolution Representation and Fuzzy Logic', *International Journal of Computer Science and Information Technology*, vol. 4, no. 1, pp. 117–126, 2012.
- [6] K. Wang, G. Lavoué, F. Denis, and A. Baskurt, 'Robust and blind mesh watermarking based on volume moments', *Computers & Graphics*, vol. 35, no. 1, pp. 1–19, Jan. 2011.
- [7] R. Ohbuchi, S. Takahashi, T. Miyazawa, and A. Mukaiyama, "Watermarking 3D polygonal meshes in the mesh spectral domain," in *Proc. of Graphics Interface*, Ottawa, Canada, pp. 9–17. 2001
- [8] S. Katz, G. Leifman, and A. Tal, 'Mesh segmentation using feature point and core extraction', *The Visual Computer*, vol. 21, no. 8–10, pp. 649–658, Jan. 2005.
- [9] F. Dunn, I. Parberry, "3D Math Premier for Graphics and Game Development", 1st, ed. Wordwar Publishing, Inc. 2002.
- [10] Cignoni, Rocchini, and Scopigno, 'Metro: Measuring Error on Simplified Surfaces', *Computer Graphics Forum*, vol. 17, no. 2, pp. 167–174, Jan. 1998.
- [11] K. Wang, G. Lavoué, F. Denis, A. Baskurt, and X. He, 'A Benchmark for 3D Mesh Watermarking', *2010 Shape Modeling International Conference*, pp. 231 - 235. 2010.
- [12] K. Wang, G. Lavoué, F. Denis, and A. Baskurt, "Robust and blind watermarking of polygonal meshes based on volume moments," *Journal of Computers and Graphics*, vol. 35, pp. 1-19 , 2011 .
- [13] J. W. Cho, R. Prost, and H. Y. Jung, "An oblivious watermarking for 3D polygonal meshes using distribution of vertex norms", *IEEE Trans. on Signal Process.*, vol. 55, no. 1, pp. 142-155, 2007.
- [14] R. K. Megalingam, M. M. Nair, R. Srikumar, V. K. Balasubramanian, and V. S. V. Sarma, 'A Comparative Study on Performance of Novel, Robust Spatial Domain Digital Image

User interfaces applied to teleoperate mobile robots with keyboard command, PS3 controller and mobile phone

Nancy Velasco E, Darío José Mendoza Chipantasi, Antonio Barrientos Cruz.

Centro de Automática y Robótica
Universidad Politécnica de Madrid
Madrid, España

{ndr.velasco, dario.mendoza.chipantasi}@alumnos.upm.es, antonio.barrientos@upm.es

Abstract— User interfaces are part of our daily life, it is usual that people are connected to the digital world through applications in a computer, smart phone or tablet. At present time these devices offer a fluency feeling in real time communications, this is the principal argument to use them in this project. We present the design of three types of user interfaces employing Matlab and Android, which constitute a bridge between the movements that the person desires and mobile robot with Bluetooth connectivity. For this objective we control robots by the inclination of phone without physical buttons, also for many people who like the videogames is possible to control it with a traditional controller and finally if not available any of the above, the robot can be controlled through a common keyboard.. In addition this paper show the design and control of two cheap mobile robots for the user interface demonstration which contain an Arduino like a main processing board and the case created with 3d printer.

Keywords— *mobile robots, teleoperate, multimedia interfaces, PS3 controller.*

I. INTRODUCTION

Nowadays user interfaces development are becoming increasingly crucial. One of the most significant types of development is based in Android, the operating system that powers millions smartphones and tablets, fast and smooth with slick graphics [1].

The future is wireless and Bluetooth technology [2] is a favorite in the world of electronics enthusiasts where the data link "no physical connection" must be robust, reliable and secure. Moreover the operative system Android [3] is the most famous in the world [4]. Furthermore, in recent years research on mobile robots is gaining followers [5][6].

An action field of mobile robots is the teleoperation [7], teleoperation allows an operator in a specific place to execute a task on another place, possibly separated by large distances [8][9]. On other hand, game controllers are used for different objectives like [10][11], one of them is the PlayStation controller that contains enough buttons to program robot functions like the velocity, steering and connectivity.

In this paper we combine the potential of android, ease of Bluetooth wireless connection, focusing on design a small cheap mobile robot [12] teleoperated via a smartphone or pc.

The robot has an Arduino control board, a Bluetooth module for wireless communication, infrared or ultrasonic sensors (depending on the model) to avoid collisions.

The control allows to link the robot with the device operated by the user with different speeds and automatic avoidance of obstacles, the control stops the movements of robot with the exception of reversing.

The rest of this paper is organized as follows. Section II: gives a simple summary about the user interfaces developed for control the mobile robots using keyboard, ps3 controller and mobile phone. Section III presents the design of robots, information about theirs parts and performance, as well as design drawings with their respective views. The experiments are given in Section IV. Finally, Section V is the conclusion.

II. USER INTERFACES

A. *The heart of user interfaces*

The user interfaces are the skin of applications they represent the link between the codes that be sending via Bluetooth and the friendly user interfaces. Always the best program or application need be simple and intuitive; for these reasons we designed three different interfaces, but the heart of communication of them are the same.

These robots used Bluetooth technology that is mainly based on connect two devices together with the same settings. A brief explanation about the parameters for serial communication is listed below:

Baud Rate = 9600;

Byte Size = 8;

Stop Bits = ONE STOP BIT;

Parity = NON PARITY;

Once configured these parameters on both mobile robot and control device, the robots are ready to begin the transmission of data. We devise a common communication for send the same data although we change the user interface,

The possible movements are forward, backward, left and right, the speeds that can be chosen are: fast, normal and slow, we contemplated that if the user does not press the scroll keys, the robot will stop.

B. PS3 Controller Interface

The DUALSHOCK 3 wireless controller for the PlayStation 3 system provides the most intuitive game play experience with pressure sensors in each action button and the inclusion of the highly sensitive SIXAXIS technology motion detection. Each hit, crash and explosion is more realistic when the user feels the rumble right in the palm of your hand. Use Bluetooth wireless technology for gaming [13].

An user interface Fig. 1 was created in Matlab to control robots using the PS3 controller, this program allows you perform movements such as forward, backward, left and right, as they have pressed the corresponding keys on the controller, plus you can control the speed of moving in three distinct

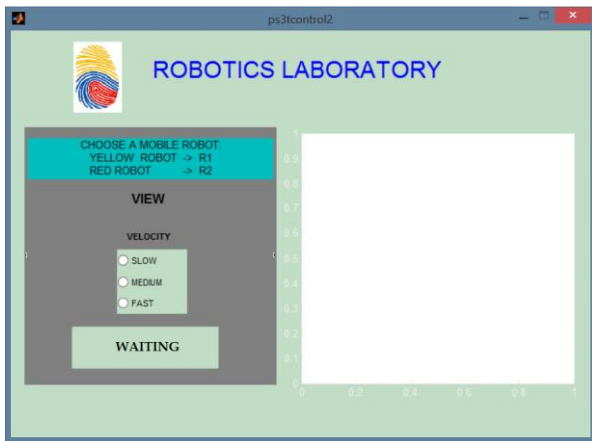


Fig. 1. Interface for ps3 controller

stages: the speeds that can be selected are low, medium, high.

First we set the control to work in Matlab with mfiles of JOYMEX [14]. These platforms lead the development of this work;

We use mfiles of JOYMEX because it interprets the signals sent by each button of ps3 controller, for this way we can change or finish the control. Although we do not use in this project, ps3 controller has there buttons that give pressure information on them and even gives the rotation of the ps3 controller, for our project we focus on recognizing the data from the buttons are explained below;

i. Selecting the robot

- Yellow mobile robot- must press the "R1" button.
- Red mobile robot - must press the "R2" button.

ii. After selecting the robot with which we interact, we must expect the GUIDE notify us that found within the range of the

robot and start connecting calls via Bluetooth, for pairing must press the "L1" button and wait the next step.

iii. Selecting the moving speed of the robot:

- Slow - button "triangle".
- Medium - button "circle".
- Fast - button "X".

iv. Controlling the robot with the arrow keys on the remote:

- Forward
- Right
- Backward
- Left

v. terminate the application press the "START" RC button

C. Keyboard Interface

In the application Fig. 2 the mobile robot is selected (yellow or red) and also the speed of movement, when you click on the start button, the application sets the Bluetooth communication, the status indicator will change from offline to online. The control keys of mobile robot are:

- w: forward
- d: right
- s: backward
- a: left

The application is terminated with finish button.

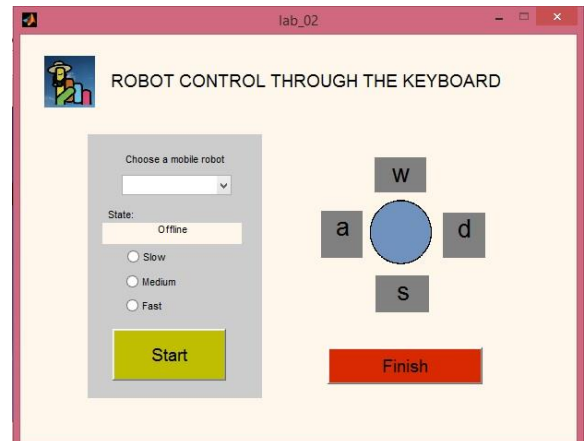


Fig. 2. Keyboard interface.

D. Phone Interface

The application Fig. 3 has the following elements:

- Power button, when pressed and select the device to be connected, the phone can be used as a remote control.
- A connection status indicator, it is connected (green) and offline (red).
- Four indicators are showing the move instruction, (left, right, up, back) and speed (yellow = slow,

orange = Mede, and red = fast). The control is sensitive to the phone inclination, if the phone inclination increase, the velocity will bigger, but when the phone is centered the indicators are gray and the robot is stopped.

- Shutdown button to terminate the application.

The application has designed similar to following images:

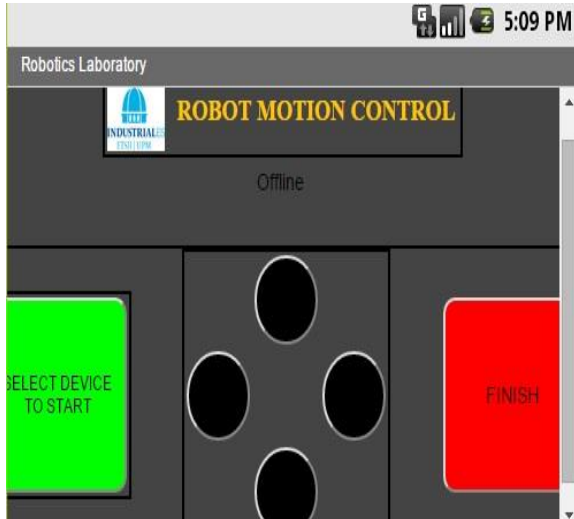


Fig. 3. Phone Interface

III. MOBILE ROBOTS DESIGN

We design and use two robots for the user interface demonstration.

The idea was create a cheap robot mobile, we decided to make our own designs and printed them, for this we use Inventor software, the design has complex forms not only because they look better aesthetically, but also for its functionality, we place two sensors at 45 degrees with respect at frontal sensor to avoid collisions, in the lower front the robot has an proximity infrared sensor to prevent the robot falling off the table or even for use it as a line follower.

Robots assembled and painted are in Fig. 4.



Fig. 4. Mobile Robots Design

A. General Recognition of the Robots

The views of mobile robots are in Fig. 5. Parts shown include:

1. On-off switch Battery.
2. Infrared Proximity Sensors
3. Ultrasonic Sensors
4. Floor sensor
5. Holes for jockey wheel
6. Motors
7. Lithium Polymer (LIPO).
8. Area circuitry.

Mobile robots have Arduino boards that are open platforms for prototyping with flexible software, inexpensive and easy to use.

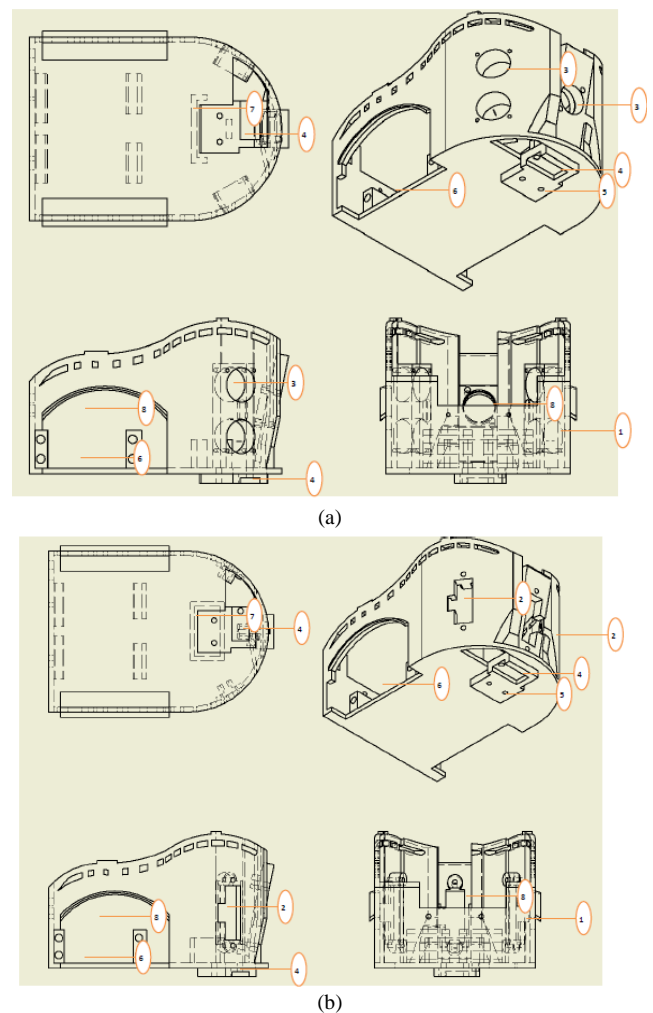


Fig.5. Views of mobile robot. (a) Robot with Ultrasonic Sensors and (b) Robot Infrared Proximity Sensors

IV. EXPERIMENTS

In our work, we designed, printed and developed two mobile robots that are controlled independently with multiple devices. We try to prove that it is easy to merge several technologies to make interesting projects with inexpensive cost and with customizable user interfaces, for us the most innovative controller was the smartphone because it is able to control robots by the inclination without physical buttons, but for many people who like the videogames is also possible to control it with a traditional controller and finally if not available any of the above, the robot can be controlled through a common keyboard.

Experiments with the ps3 controller allowed to conclude that the interface was adequate for handling the mobile robot, since many users are familiarized to the control because it is a command for video games. Fig. 6 shows the application that run and the user with ps3 controller.

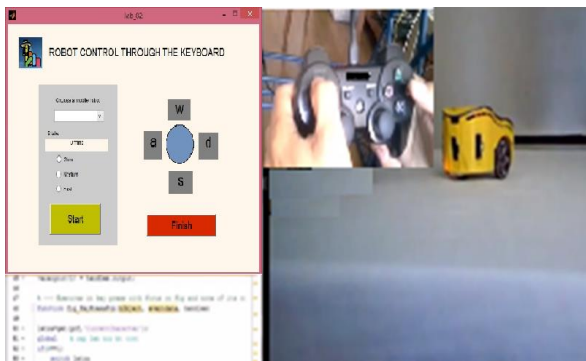


Fig. 6. Application running with the robot responding to command

Experiments with the keyboard were efficient, each person that operate a computer has used a keyboard, so the interface is simple to use. The Fig. 7. Shows the application and the robot.

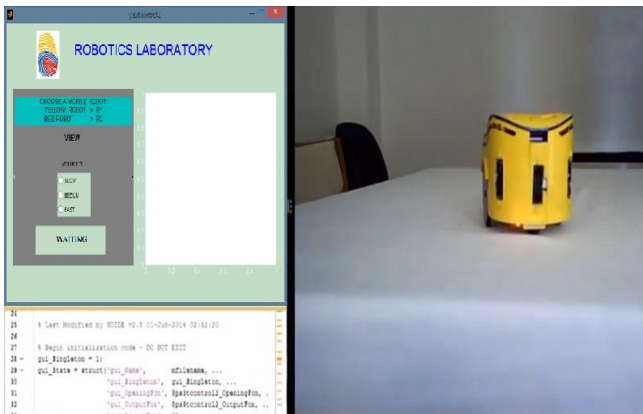


Fig. 7. Keyboard Application and the robots responding to command.

Experiments with the phone were the most intuitive control using the accelerometer [15] as sensor, thereby moving the phone in one direction and with a certain inclination user can

generate a slow or fast motion or the detention of the robot

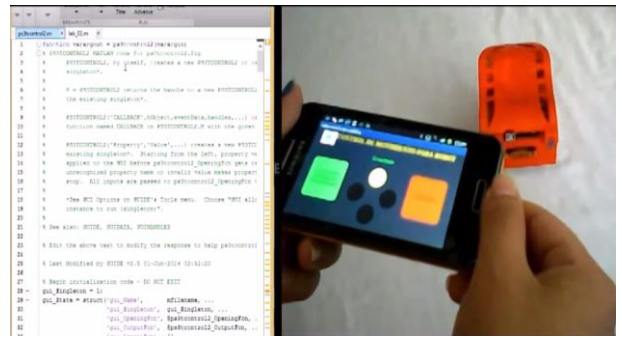


Fig. 8. Phone Application and robot responding to command.

when is centered. Fig. 8. Shows this action.

ACKNOWLEDGMENT

Extensive gratitude to Professor Antonio Barrientos for having proposed the development of this challenge. Nancy Velasco and Darío Mendoza are supported by the Secretaría Nacional de Educación Superior, Ciencia, Tecnología e Innovación SENESCYT (Quito, Ecuador) under Convocatoria Abierta 2013 Scholarship Program.

REFERENCES

- [1] Lu Ying; Tang Xiao-jun; Liu Na; Mao Yu-Yue; Li Ming-Xia; Xiao Peng; Wang Hai-Wen, "Application and Research of Mobile Terminal with Android in the Equipment Monitoring System," Parallel Architectures, Algorithms and Programming (PAAP), 2014 Sixth International Symposium on , vol., no., pp.293,296, 13-15 July 2014
- [2] Davies, A.C., "An overview of Bluetooth Wireless Technology™ and some competing LAN standards," Circuits and Systems for Communications, 2002. Proceedings. ICCSC '02. 1st IEEE International Conference on , vol., no., pp.206,211, 2002.
- [3] Djajadi, A.; Putra, R.J., "Inter-cars safety communication system based on Android smartphone," Open Systems (ICOS), 2014 IEEE Conference on , vol., no., pp.12,17, 26-28 Oct. 2014.
- [4] <https://developer.android.com/about/index.html>
- [5] Ghayas, S.; Sulairnan, S.; Jaafar, J.; Mahammad, S., "Motivational scaffolding: Tackling the challenges of using mobile applications," User Science and Engineering (i-USER), 2014 3rd International Conference on , vol., no., pp.48,51, 2-5 Sept. 2014.
- [6] Vithani, T.; Kumar, A., "Presentation 5. A comprehensive mobile application development and testing lifecycle," IT Professional Conference (IT Pro), 2014 , vol., no., pp.1,3, 22-22 May 2014.
- [7] Jianping Cai; Jianzhong Wu; Minghui Wu; Meimei Huo, "A bluetooth toy car control realization by android equipment," Transportation, Mechanical, and Electrical Engineering (TMEE), 2011 International Conference on , vol., no., pp.2429,2432, 16-18 Dec. 2011
- [8] C. Sayers, Remote Control Robotics, Springer-Verlag, 1998.
- [9] Alencastre-Miranda, M.; Munoz-Gomez, L.; Rudomin, I., "Teleoperating robots in multiuser virtual environments," Computer

Science, 2003. ENC 2003. Proceedings of the Fourth Mexican International Conference on , vol., no., pp.314,321, 8-12 Sept. 2003.

- [10] Nagata, F.; Watanabe, K., "Teaching system for a polishing robot using a game joystick," SICE 2000. Proceedings of the 39th SICE Annual Conference. International Session Papers , vol., no., pp.179,184, 2000.
- [11] Ching-Chang Wong; Wei-Wen Wang; Ya-Ling Lee; Cheng-Hui Wu, "Remote controlled game platform by USB joysticks," Mechatronics, 2005. ICM '05. IEEE International Conference on , vol., no., pp.136,139, 10-12 July 2005.
- [12] A. Barrientos, Fundamentos de Robótica, 2nd ed. McGraw-Hill, 2007.
- [13] <http://us.playstation.com/ps3/accessories/dualshock-3-wireless-controller-ps3.html>
- [14] <http://joymex.escabe.org/>
- [15] <http://www.androidpit.es/funcionamiento-acelerometro-smartphones>

vowels as HCI for Controlling Mouse Cursor

Mohamed FEZARI, NouhaBounouioua

Badji Mokhtar Annaba University, Faculty of Engineering, BP: 12, Annaba, 23000,
Annaba, Algeria
mohamed.fezari@univ-annaba.dz, n.bounoui@yahoo.fr

Ahmed Al-Dahoud

University of Bradford,
Bradford, UK

ahmad_aldahoud@hotmail.com,

Abstract— HCI (Human Computer Interfaces) applications are generally based on using keyboard or joystick; in this paper we experimented the use of vowels to activate some input device such as to control the movement of mouse pointer on the screen. The control of the windows icon mouse pointer (WIMP) by voice command is currently based on using vowel utterances, this category of letters is easy to recognize and to be pronounced, especially for individuals who are physically disabled or have a partial voice disorder. So this type of MCI might be used by a category of disabled person. In addition, vowels are quite easy to model by automatic speech recognition (ASR) systems. In this work we represent the design of a system for the control of mouse cursor based on voice command, using the pronunciation of certain vowels and short words. The Mel Frequency Cepstral Coefficients (MFCCs), fundamental frequency (F_0) and Formants (F_1, F_2) are selected as features. The TDW with Euclidian Distance and Hidden Markov Models (HMMs) have been tested as classifiers for matching components (vowels and short words). Comparison between different features and classifiers were tested and results are presented on tables, finally a GUI has been designed for user applications.

Keywords- vowel recognition; dynamic time warping; MFCC features; HMM.

I. INTRODUCTION

Existing human-computer interfaces are not suited to individuals with upper limb motor impairments. Recently, a lot of interest is put on improving all aspects of the interaction between human and computer especially for this category of persons, however these devices are generally more expensive example sip-and-switches [1] eye-gas and eye tracking devices[4], head mice [2,3] chin joystick[5] and tongue switches [6]. Here is some related works on human computer interaction, based on voice activation or control, which can be invested for individuals with motor impairments. Most of concepts of vocal commands are built on the pronunciation of vowels [5, 6, and 7], where the particularity of vowels used is the simple and the regular pronunciation of these phonemes. Many vocal characteristics are exploited in several works, but the most used are: energy [1, 2, 3 and 5], pitch and vowel quality [9,10] speech rate (number of syllables per second) and volume level [7]. However, Mel Frequency Cepstral Coefficients (MFCCs) [11, 12 and 13] are used significantly of speech processing as bio-inspired feature for automatic speech recognition of isolated words [15-16].

The paper is organized as follows: in section 2, presentation of an overview on related works of mouse cursor control based

on voice control and commands. In section 3, we showed LPC and MFCC computation and use as features extraction techniques. Then we describe used classifiers: DTW then HMM in section 4. In section 5, we present tests and results. And finally, we provide graphic user interface as an application.

II. RELATED WORKS:

We describe some related works with vocal command system in the literature review. Voice recognition allows you to provide input to an application with your voice. In the basic protocol, each vowel is associated to one direction for pointer motion [1]. This technique is useful in situations where the user cannot use his or her hands for controlling applications because of permanent physical disability or temporal task-induced disability. The limitation of this technique is that it requires an unnatural way of using the voice [5] [6]. Control by Continuous Voice: In this interface, the user's voice works as an on/off button. When the user is continuously producing vocal sound, the system responds as if the button is being pressed. When the user stops the sound, the system recognizes that the button is released. For example, one can say "Volume up, ahhhhhh", and the volume of a TV set continues to increase while the "ahhh" continues. The advantage of this technique compared with traditional approach of saying "Volume up twenty" or something is that the user can

continuously observes the immediate feedback during the interaction. One can also use voiceless, breathed sound [6].

Alex Olwal et al. [7] have been experimenting with non verbal features in a prototype system in which the cursor speed and direction are controlled by speech commands. In one approach, speech commands provide the direction (right, left, up and down) and speech rate controls the cursor speed. Mapping speech rate to cursor speed is easy to understand and allows the user to execute slow. The cursor's speed can be changed while it is moving, by reissuing the command at a different pace. One limitation of using speech features is that they are normally used to convey emotion, rather than for interaction control.

The detection of gestures is based on discrete pre-designated symbol sets, which are manually labeled during the training phase. The gesture-speech correlation is modeled by examining the co-occurring speech and gesture patterns. This correlation can be used to fuse gesture and speech modalities for edutainment applications (i.e. video games, 3-D animations) where natural gestures of talking avatars are animated from speech [7] [8].

J. Bilmes et al. [9] have been developed a portable modular library (the Vocal Joystick"VJ" engine) that can be incorporated into a variety of applications such as mouse and menu control, or robotic arm manipulation. Our design goal is to be modular, low-latency, and as computationally efficient as possible. The first of those, localized acoustic energy is used for voice activity detection, and it is normalized relatively to the current detected vowel, and is used by our mouse application to control the velocity of cursor movement. The second parameter, "pitch", is not used currently but it is left for the future use. The third parameter: "vowel quality", where the vowels are characterized by high energetic level. The classification of vowels is realized by extraction of two first formants frequencies, tongue height and tongue advancement [9, 10]. Thus, the VJ research has focused on real time extraction of continuous parameters since that is less like standard ASR technology [9]. The main advantage of VJ is the reaction of the system in real time.

In [14], Thiang et al., described the implementation of speech recognition system on a mobile robot for controlling movement of the robot. The methods used for speech recognition system are Linear Predictive Coding (LPC) and Artificial Neural Network (ANN). LPC method is used for extracting feature of a voice signal and ANN is used as the recognition method. Backpropagation method is used to train the ANN. Experimental results show that the highest recognition rate that can be achieved by this system is 91.4%. This result is obtained by using 25 samples per word, 1 hidden layer, 5 neurons for each hidden layer, and learning rate 0.1.

III. FEATURE EXTRACTION

In order to implement the HMI application on embedded system in future, and to get good results in automatic speech recognition is to select better and easy to compute features, so the

features would be robust and fast to compute. The LPC, MFCC with energy and derivatives were selected based on literature reviews [15, 16] and [17].

A. MFCC Feature extraction[11]

The extraction of the best parametric representation of acoustic signals is an important task to produce a better recognition performance. The efficiency of this phase is important for the next phase since it affects its behavior. MFCC is based on human hearing perceptions which cannot perceive frequencies over 1Khz. In other words, in MFCC is based on known variation of the human ear's critical bandwidth with frequency. MFCC has two types of filter which are spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000Hz. A subjective pitch is present on Mel Frequency Scale to capture important characteristic of phonetic in speech. The overall process of the MFCC can be presented in the following steps:

1. After the pre-emphasis filter, the speech signal is first divided into fixed-size windows distributed uniformly along the signal.
2. The FFT (Fast Fourier Transform) of the frame is calculated. Then the energy is calculated by squaring the value of the FFT. The energy is then passed through each filter Mel. S_k : is the energy of the signal at the output of the filter K, we have now m_p (number of filters) S_k parameters.
3. The logarithm of S_k is calculated.
4. Finally, the coefficients are calculated using the DCT (Discrete Cosine Transform).

$$c_i = \sqrt{\frac{2}{m_p}} \left\{ \sum_{k=1}^{m_p} \log(S_k) \cos \left[i \left(k - \frac{1}{2} \right) \frac{\pi}{m_p} \right] \right\} \quad (1)$$

pour $i = 1 \dots \dots N$

N: is the number of MFCC coefficients.

B. Fundamental Frequency and formants extraction

Linear predictive analysis of speech has become the predominant technique for estimating the basic parameters of speech. Linear predictive analysis provides both an accurate estimate of the speech parameters and also an efficient computational model of speech.

The basic idea behind linear predictive analysis is that a specific speech sample at the current time can be approximated as a linear combination of past speech samples. Through minimizing the sum of squared differences (over a finite interval) between the actual speech samples and linear predicted values a unique set of parameters or predictor coefficients can be determined.

LPC computation basic steps can be presented as follow [14]:

a) *Pre-emphasis*: The digitized speech signal, $s(n)$, is put through a low order digital system, to spectrally flatten the signal and to make it less susceptible to finite precision effects later in the signal processing.

b) *Frame Blocking*: The output of pre-emphasis step $\tilde{s}(n)$ is blocked into frames of N samples, with adjacent frames being separated by M samples. If $x_l(n)$ is the l^{th} frame of speech, and there are L frames within entire speech signal.

c) *Windowing*: After frame blocking, the next step is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. If we define the window as $w(n)$, $0 \leq n \leq N-1$, then the result of windowing is the signal:

$$\tilde{x}_l(n) = x_l(n)w(n) \quad (2)$$

d) *Autocorrelation Analysis*: The next step is to auto correlate each frame of windowed signal in order to give:

$$r_l(m) = \sum_{n=0}^{N-1-m} \tilde{x}_l(n)\tilde{x}_l(n+m) \quad (3)$$

$$m = 0, 1, \dots, p$$

e) *LPC Analysis*: which converts each frame of $p + 1$ autocorrelations into LPC parameter set by using Durbin's method.

f) *LPC Parameter Conversion to Cepstral Coefficients*: LPC cepstral coefficients, is a very important LPC parameter set, which can be derived directly from the LPC coefficient set. The recursion used is:

$$c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) \cdot c_k \cdot a_{m-k} \quad 1 \leq m \leq p \quad (4)$$

And:

$$c_m = \sum_{k=m-p}^{m-1} \left(\frac{k}{m}\right) \cdot c_k \cdot a_{m-k} \quad (5)$$

$$m > p$$

The LPC cepstral coefficients are the features that are extracted from voice signal and these coefficients are used as the input data for the classifier (Euclidian Distance or DTW). In this system, voice signal is sampled using sampling frequency of 8 kHz and the signal is sampled within 1.5 seconds, therefore, the sampling process results 1200 data. Because we choose LPC parameter $N = 200$, $m = 100$, and LPC order = 10 then there are 119 vector data of LPC cepstral coefficients.

IV. CLASSIFIERS

In pattern recognition in general, automatic speech recognition, speaker Identification, image or shape recognition we need some how an algorithm to classify.

A. DTW(Dynamic Time Warping)

DTW algorithm is based on Dynamic Programming techniques .This algorithm is for measuring similarity between two time series which may vary in time or speed. This technique also used to find the optimal alignment between two times series if one time series may be "warped" non-linearly by stretching or shrinking it along its time axis.

This warping between two time series can then be used to find corresponding regions between the two time series or to determine the similarity between the two time series.

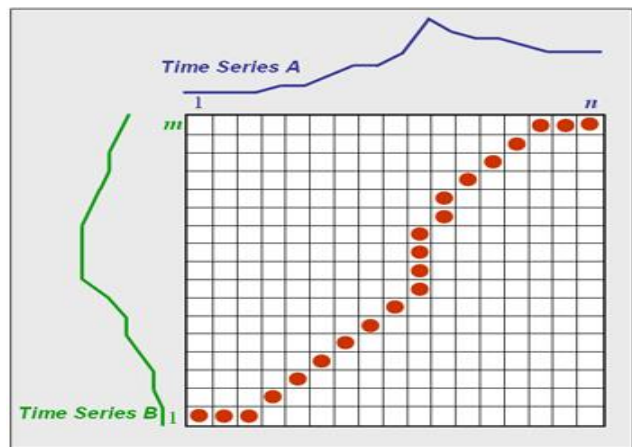


Fig. 1. The optimal warping path from [22]

B. Euclidian distance formulat:

The Euclidean distance between points p and q is the length of the line segment connecting them (\overline{PQ}).

In Cartesian coordinates, if $\mathbf{p} = (p_1, p_2, \dots, p_n)$ and $\mathbf{q} = (q_1, q_2, \dots, q_n)$ are two points in Euclidean n -space, then the distance (d) from \mathbf{p} to \mathbf{q} , or from \mathbf{q} to \mathbf{p} is given by the Pythagorean formula:

$$\text{dist}((x, y), (a, b)) = ((x - a)^2 + (y - b)^2)^{1/2} \quad (6)$$

$$\text{Dist}(q_i, p_i) = \text{Sum}(q_i - p_i)^2 \quad \text{for } i=1..n \quad (7)$$

The position of a point in a Euclidean n -space is a Euclidean vector. So, \mathbf{p} and \mathbf{q} are Euclidean vectors, starting from the origin of the space, and their tips indicate two points. The **Euclidean norm**, or **Euclidean length**, or **magnitude** of a vector measures the length of the vector:

$$\|\mathbf{p}\| = \sqrt{p_1^2 + p_2^2 + \dots + p_n^2} = \sqrt{\mathbf{p} \cdot \mathbf{p}} \quad (8)$$

where the last equation involves the dot product.

A vector can be described as a directed line segment from the origin of the Euclidean space (vector tail), to a point in that space (vector tip). If we consider that its length is actually the distance from its tail to its tip, it becomes clear that the Euclidean norm of a vector is just a special case of Euclidean distance: the Euclidean distance between its tail and its tip.

The distance between points \mathbf{p} and \mathbf{q} may have a direction (e.g. from \mathbf{p} to \mathbf{q}), so it may be represented by another vector, given by

$$\mathbf{q} - \mathbf{p} = (q_1 - p_1, q_2 - p_2, \dots, q_n - p_n) \quad (9)$$

If $D(x,y)$ is the Euclidean distance between frame x of the speech sample and frame y of the reference template, and if $C(x,y)$ is the cumulative score along an optimal alignment path that leads to (x,y) , then:

$$C(x,y) = \min(C(x-1,y), C(x-1,y-1), C(x,y-1)) + D(x,y) \quad (10)$$

C. HMMs Basics [13]

Over the past years, Hidden Markov Models have been widely applied in several models like pattern, or speech recognition. To use a HMM, we need a training phase and a test phase. For the training stage, we usually work with the Baum-Welch algorithm to estimate the parameters (π, A, B) for the HMM. This method is based on the maximum likelihood criterion. To compute the most probable state sequence, the Viterbi algorithm is the most suitable.

An HMM model is basically a stochastic finite state automaton, which generates an observation string, that is, the sequence of observation vectors, $O = O_1, \dots, O_t, \dots, O_T$. Thus, a HMM model consists of a number of N states $S = \{S_i\}$ and of the observation string produced as a result of emitting a vector 'Ot' for each successive transitions from one state S_i to a state S_j . 'Ot' is d dimension and in the discrete case takes its values in a library of M symbols.

The state transition probability distribution between state S_i to S_j is $A = \{a_{ij}\}$, and the observation probability distribution of emitting any vector 'Ot' at state S_j is given by $B = \{b_j(O_t)\}$. The probability distribution of initial state is $\Pi = \{\pi_i\}$.

$$a_{ij} = P(q_{t+1} = S_j | q_t = S_i) \quad (11)$$

$$B = \{b_j(O_t)\} \quad (12)$$

$$\pi_i = P(q_0 = S_i) \quad (13)$$

Given an observation O and a HMM model $\lambda = (A, B, \Pi)$, the probability of the observed sequence by the forward-backward

procedure $P(O/\lambda)$ can be computed. Consequently, the forward variable is defined as the probability of the partial observation sequence O_1, O_2, \dots, O_t (until time t) and the state S at time t , with the model λ as $\alpha(i)$. and the backward variable is defined as the probability of the partial observation sequence from $t+1$ to the end, given state S at time t and the model λ as $\beta(i)$. The probability of the observation sequence is computed as follow:

$$p(o/\lambda) = \sum_{i=1}^N \alpha_t(i) * \beta_t(i) = \sum_{i=1}^N \alpha_T(i) \quad (14)$$

And the probability of being in state I at time t , given the observation sequence O and the model λ is computed as in (13).

V. DESCRIPTION OF APPLICATION

The application is designed to control the mouse cursor by using the pronunciation of certain phonemes and words, which we chose as vocabulary: "aaa", "ooh", "iii", "eeu", "ou", "uu", "Clic" and "stop".

The choice of these vowels and short words is based on the following criteria:

- Easy to learn.
- Easy to pronounce.
- can be pronounced persons with voice disorder.
- Easy to recognize by automatic speech recognition system.

A. DataBase Description

The database consists of 10 women (age 20 to 50 years), 10 men (age 20 to 60 years), and 5 children (age from 5 to 14 years) and category of persons with voice disorder from German database of the PTSD Putzer's voice in [18], each speaker had: 5 trials for each phoneme or word. Collection of the database is performed in a quiet room without noise.

B. The parametrization

According to the tests, we found that the parameters more robust to noise than other parameters are the LPC coefficients and Mel Frequency Cepstral Coefficients (MFCCs). The input signal is segmented by a window of 25 ms overlapping 10ms, from each segment parameters were extracted by both methods LPC (the order of the prediction: 10) then MFCC (42 coefficients: Energy and derivative and second derivatives).

C. Classification

For this moment, we have tested two classifier, first one has been used for simplicity in order to be implemented in future on

DSP circuit of microcontroller: Dynamic Time Warping (DTW) with Euclidian distance and Hidden Markov chains (HMM) for classification phase.

For Hidden Markov models, in our system, we utilize left-to-right HMM structures with 3 states and 3 mixtures are used to model MFCCs coefficients.

D. Application

Our application is used to control the mouse cursor by voice, pronouncing a vowel or short words above. The vowels are mapped to directions of movement cursor and push buttons on mouse as follow and presented in figure 2:

- Up:" ooh"
- Down:" aah"
- To the right:" iii"
- Left:" eeu"
- To double-click (open):" click" or "eke"
- To exit the application by voice command:" stop" or"abe"
- Left-Click : "ou"
- Right-Click:" "uu"

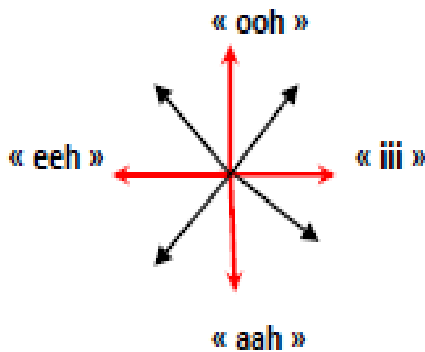


Fig. 2. Directions of cursor mouse mapping from vowels

VI. RESULTS AND DISCUSSIONS

For the testing phase, 20% of recorded sounds are selected for each vowel or short word from the vocabulary.

In order to see the effect of training and making the system speaker independent, different scenarios for the tests were done, where we choose the results of recognition of three users out of database.

Some vowels and short words were correctly classified with some confusion, where a phoneme (or word) test classified as another phoneme (or word), the misclassification is presented in the tables below (I, II). And it is clear that the confusion is higher in LPC features with DTW classifier while it is reduced using MFCC with HMM classifier.

TABLE I. CONFUSION TABLE USING (MFCC/HMM)

Pronounced Vowel	Classified as:						
	aaa	ooh	eeu	iii	clic	stop	ou
aaa	o	x	-	-	-	x	-
ooh	x	o	-	-	-	x	-
eeu	-	x	o	x	-	-	x
iii	-	-	-	o	x	-	-
ou	-	x	x	-	-	-	o
uu	-	-	x	-	-	-	x
Clic or "eke"	-	-	-	-	o	x	-
Stop or "ebe"	-	x	-	-	-	o	-

x: means that pronounced phoneme classified as an other

TABLE II. CONFUSION TABLE USING (LPC/DTW)

Pronounced Vowel	Classified as:						
	aaa	ooh	eeu	iii	clic	stop	ou
aaa	o	x	-	-	x	x	-
ooh	x	o	x	-	-	-	x
eeu	x	x	o	x	-	-	x
iii	x	-	x	o	x	-	-
ou	-	x	-	-	-	x	o
uu	-	x	x	-	-	-	x
Clic or "eke"	-	-	x	x	o	-	-
Stop or "ebe"	-	x	x	-	x	o	x

x: means that pronounced phoneme classified as an other

TABLE III. CLASSIFICATION USING LPCS, MFCC AND DTW AS CLASSIFIER

Vowel	LPC (%)	MFCC (%)
aaa	76	81
ooh	58.33	62
eeu	57	59
iii	61	73
Clic or "eke"	54.55	79
Stop or "ebe"	55.56	81

TABLE IV. CLASSIFICATION USING LPC, MFCCS AND HMM AS CLASSIFIER

Vowel	LPC (%)	MFCC (%)
aaa	85	92
ooh	78	83
eeu	74	84
iii	79	87
clic	87	94
stop	90	95

According to the results presented above (Tables: III, IV), the recognition rates using MFCCs parameterization classification with DTW or HMM classification is better than: LPCs and MFCCs with DTW. So we can say that the MFCCs / HMM system is partially independent of the speaker.

Results using MFCC and HMM, on German database vowels (sounds) for persons with chronic inflammation of the larynx and vocal fold nodules [19], are presented in Table V.

TABLE V. CLASSIFICATION USING LPC AND MFCC USING HMM FOR VOWEL FORM GERMAN DB [18]

Vowel	LPC (%)	MFCC (%)
aaa	55	67
ooh	42	53
eeu	43	49
iii	53	72
Clic or "eke"	57	64
Stop or "ebe"	54	70

We can see that the recognition rate is little bit lower than for healthy persons, we conclude that in this case other special features might be necessary to include on the application.

In addition, we must consider the preprocessing for noise in future work, as well as the database training models need from the category of children.

CONCLUSION

According to the results, we note that the classification using HMM is better than the DTW, and the decision based on MFCC coefficients is more certain than the coefficients LPCs.

From experimental results, it can be concluded that MFCC features and HMM as classifier can recognize the speech signal well. Where the highest recognition rate that can be achieved in the last scenario. This result is achieved by using MFCCs and HMM. Moreover, we need to get better features to improve classification of vowel and short words pronounced from voice disabled persons; in fact this can be resolved by inserting Jitter and Shimmer as features.

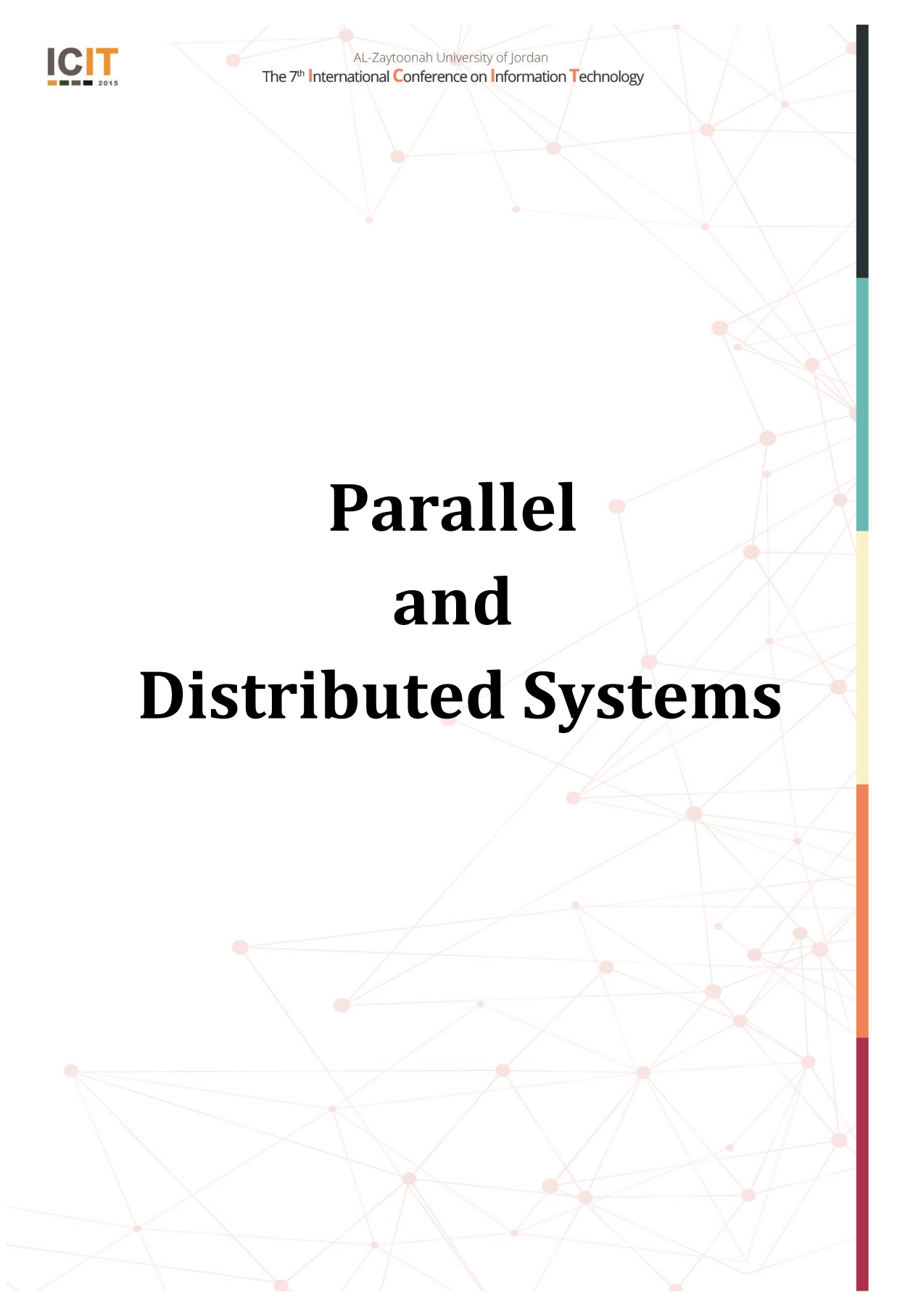
We notified that the variety of signals, collected for database from different age and gender, the recording conditions and the environment, have a considerable impact in classification results.

REFERENCES

- [1] "Pride mobility products group sip-n-puff system/head array control", 2005, <http://pridemobility.com>
- [2] "origine instruments sip/puff switch and head mouse", 2005, orin.com/access/headmouse/index.com
- [3] "Headmaster head mouse", 2003, <http://wati.com/headmaster.htm>
- [4] "assistive technologies's eye gaze system for computer access", 2003, <http://www.assistivetechologies.com/proddetails/EG001B.htm>
- [5] C. de Mauro, M. Gori, M. Maggini, and E. Martinelli, "A voice device with an application-adapted protocol for Microsoft windows," In Proc. IEEE Int. Conf. on Multimedia Comp. and Systems, vol. 2, pp. 1015–1016, Firenze, Italy, 1999.
- [6] T. Igarashi and J. F. Hughes, "Voice as sound: Using non-verbal voice input for interactive control," In ACM UIST 2001, November.
- [7] Alex Olwal and Steven Feiner, "Interaction techniques using prosodic features of speech and audio localization," In IUI '05: Proc. 10th Int. Conf. on Intelligent User Interfaces. New York: NY, USA, 2005. ACM Press, pp. 284–286.
- [8] M.E. Sargin, O. Aran, A. Karpov, F. Oflil, Y. Yasinnik, S. Wilson, E. Erzin, Y. Yemez and A.M. Tekalp, "Combined Gesture-Speech Analysis and Speech Driven Gesture Synthesis," ICME 2006 : IEEE International Conference on Multimedia and Expo, July 2006, pp: 893–896.
- [9] J. Bilmes, X. Li, J. Malkin, K. Kilanski, R. Wright, K. Kirchoff, A. Subramanya, S. Harada, J. Landay, P. Dowden, and H. Chizeck, "The vocal joystick: A voice-based human-computer interface for individuals with motor impairments," in Human Language Technology Conf. and Conf. on Empirical Methods in Natural Language Processing, Vancouver, October 2005.
- [10] S. Harada, J. Landay, J. Malkin, X. Li, J. Bilmes, "The Vocal Joystick: Evaluation of Voice-based Cursor Control Techniques", *ASSETS '06*, October 2006.
- [11] Lindsalwa Muda, Mumtaj Begam and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", *Journal of Computing*, Volume 2, Issue 3, March 2010, pp : 138-143.

- [12] Mahdi Shaneh and Azizollah Taheri, "Voice Command Recognition System Based on MFCC and VQ Algorithms" ,World Academy of Science, Engineering and Technology 57 2009, pp: 534-538.
- [13] A Bala, A Kumar, N Birla - Anjali Bala et al., "Voice command recognition system based on MFCC and DTW International Journal of Engineering Science and Technology, Vol. 2 (12), 2010, pp :7335-7342.
- [14] Thiang, S. Wijoyo, "Speech recognition using linear predictive coding and artificial neural network for controlling movement of mobile robot", International Conference on Information and Electronics Engineering IPCSIT vol.6, 2011, 179-183.
- [15] C. Snani, "conception d'un system dereconnaissance de mots isolés à base de l'approchestochastique en temps réel : Application commande vocale d'une calculatrice", Mémoire de magister ,Institut d'électronique univ. Badji mokhtar Annaba,2004.
- [16] C. HAdri, M boughazi and M fezari, "improvement of Arabic digits recognition rate based in the parameters choice", in proceedings of international conf. CISA Annaba, june 2008.
- [17] M. Fezari and A. Al-dahoud, "An Approach For: Improving Voice Command processor Based On Better Features and Classifiers Selection," pp. 1-5. The 13th International Arab Conference on Information Technology ACIT'2012 Dec.10-13 ,2012.
- [18] Manfred Putzer & Jacques Koreman " A german databse for a pattern for vacal fold vibration " Phonus 3, Institute of Phonetics, University of the Saarland, 1997, 143-153.
- [19] I.M. M. El Emary,M. Fezari, F. Amara," Towards Developing a Voice Pathologies Detection System", in Jouranal of Electronics and Communication, 2014 Elsevier.
- [20] [Eamonn J. Keogh, Michael J. Pazzani, "Derivative Dynamic Time Warping" In Proc. Of the 1st SIAM Int.Conf. on Data Mining (SDM-2001).
- [21] H. Sakoe, S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition". IEEE Transaction on Acoustics, Speech and Signal Processing, Vol 26, NO1, pp. 43-49. February 1978.
- [22] <http://cst.tu-plovdiv.bg/bi/DTWimpute/DTWalgorithm.html>

Parallel and Distributed Systems



An Effective Parallel FDTD Algorithm For Modeling 3D Frequency-Dependent Electromagnetic Applications

Omar Ramadan, Muhammed Salamah and Ahmad Salh

Computer Engineering Department
Eastern Mediterranean University
GaziMagusa, Mersin 10, Turkey

omar.ramadan@emu.edu.tr; muhammed.salamah@emu.edu.tr; asdwifi@googlemail.com

Abstract—Full-wave parallel finite difference time domain (FDTD) algorithm is presented for modeling open region dispersive electromagnetic applications. The algorithm is based on spatial partitioning of the problem geometry into adjacent non-overlapping sub-domains using two-dimensional topology. The inter-processor communication among the neighboring processors is carried out by using the message passing interface (MPI) library. The performance of the proposed parallel system, which is composed of 16 PCs interconnected through 100Mbps Ethernet, was illustrated for a point source radiating in three dimensional Lorentz dispersive domain and it has been found that the proposed algorithm not only speed up computations but also increases the maximum solvable problem size.

Keywords—Parallel programming; message passing interface (MPI); finite difference time domain (FDTD); anisotropic perfectly matched layer (APML); dispersive media.

I. INTRODUCTION

In the last decade, the finite difference time domain (FDTD) method [1] has been widely used for solving many electromagnetic problems [2]. This is due to its simplicity and direct applicability to Maxwell's curl equations. Nevertheless, when the FDTD method is used for modeling open region problems, efficient absorbing boundary conditions (ABCs) are needed to truncate the computational domains. The perfectly matched layer (PML) [3]-[6] has been shown to be one of the most effective FDTD ABCs. This ABCs surrounds the FDTD computational domain with a lossy layer that absorbs outgoing waves with minimal reflections.

To model large problems using the FDTD method, intensive computational time and memory storage are needed. Hence, parallelizing the FDTD method has been shown to be one of the latest challenges in the FDTD research. In last few years, different parallel PML-FDTD algorithms, based on the message passing interface (MPI) library [7], have been successfully introduced [8]-[14]. Nevertheless, these algorithms are suitable only for non-dispersive electromagnetic applications. In [15], dispersive parallel scalar wave equation FDTD algorithm has been presented. This approach, however, is valid only for source free applications only.

In this paper, full-wave parallel PML-FDTD algorithm is presented for modeling open region dispersive electromagnetic problems. The algorithm is based on spatial partitioning of the problem geometry into adjacent non-overlapping sub-domains using two-dimensional (2-D) topology and the inter-processor communication among the neighboring processors is carried out by using the MPI library. The performance of the proposed parallel system, which is composed of 16 PCs interconnected through 100Mbps ethernet, was illustrated for a point source radiating in three dimensional Lorentz dispersive domain and it has been observed that the parallel algorithm not only speed up computations but also increases the maximum solvable problem size. The paper is organized as follows. In Section II, the basic formulations of the FDTD and the APML ABCs approaches are presented. In Section II, the proposed parallel strategy is described. Numerical example to show the validity of the proposed parallel algorithm is included in Section III. Finally, summary and conclusions are included in Section IV.

II. THEORY

A. Basic Formulations

Considering an isotropic, homogeneous and dispersive computational domain, the frequency domain Maxwell's curl equations can be written as

$$j\omega\epsilon_0\epsilon_r(\omega)\mathbf{E}(\mathbf{r},\omega) = \nabla \times \mathbf{H}(\mathbf{r},\omega) \quad (1)$$

$$j\omega\mu_0\mathbf{H}(\mathbf{r}, \omega) = -\nabla \times \mathbf{E}(\mathbf{r}, \omega) \quad (2)$$

where \mathbf{E} and \mathbf{H} are, respectively, the electric and the magnetic field vectors, and $\varepsilon_r(\omega)$ is the relative permittivity of the domain which can be written as

$$\varepsilon_r(\omega) = \frac{\sum_{m=0}^M a_m(j\omega)^m}{\sum_{m=0}^M b_m(j\omega)^m} \quad (3)$$

where a_m and b_m , ($m = 0, 1, \dots, M$), are the coefficients of the rational polynomials and M is the maximum order of the dispersive domain. To discretize (1) and (2), consider, as an example, the E_z -field component of (1), i.e.,

$$j\omega\varepsilon_0\varepsilon_r(\omega)E_z = \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} \quad (4)$$

Equation (4) can be written as

$$j\omega\varepsilon_0 D_z = \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} \quad (5)$$

where D_z is related to E_z through the constitutive relation:

$$D_z = \varepsilon_r(\omega)E_z = \frac{\sum_{m=0}^M a_m(j\omega)^m}{\sum_{m=0}^M b_m(j\omega)^m} E_z \quad (6)$$

Using the inverse Fourier transform relation, $j\omega \Rightarrow \partial/\partial t$, and employing the FDTD time and space discretizations [1], (5) can be written in the discrete time domain as

$$D_{z,i,j,k+1/2}^{n+1} = D_{z,i,j,k+1/2}^n + \frac{\Delta t}{\Delta \varepsilon_0} \left[H_{y,i+1/2,j,k+1/2}^{n+1/2} - H_{y,i-1/2,j,k+1/2}^{n+1/2} - H_{x,i,j+1/2,k+1/2}^{n+1/2} + H_{x,i,j-1/2,k+1/2}^{n+1/2} \right] \quad (7)$$

where $\Delta = \Delta x = \Delta y$ is the space cell size. Equation (6) can be easily written in the discrete time domain easily by using the Bilinear transformation relation [16]

$$j\omega \Rightarrow \frac{z - z^{-1}}{\Delta t} \frac{1 - z^{-1}}{1 + z^{-1}} \quad (8)$$

where z^{-1} is the Z -transform variable which corresponds to a single delay element in the discrete time domain. To this end, (6) can be written in the Z -domain as

$$D_z(Z) = \frac{\sum_{m=0}^M c_m Z^{(1-m)}}{\sum_{m=0}^M d_m Z^{(1-m)}} E_z(Z) \quad (9)$$

where c_m and d_m , ($m = 0, 1, \dots, M$), are related to a_m and b_m and the time step Δt . Using the Z -transform relation

$$Z^{-m} G(Z) \rightarrow G^{n-m} \quad (10)$$

(9) can be written directly in the discrete time form as

$$E_{z,i,j,k+1/2}^{n+1} = \frac{d_0}{c_0} D_{z,i,j,k+1/2}^{n+1} + \Psi_{i,j,k+1/2}^n \quad (11)$$

where

$$\Psi_{i,j,k+1/2}^n = \frac{1}{c_0} \sum_{m=1}^M \left(d_m D_{z,i,j,k+1/2}^{n+(1-m)} - c_m E_{z,i,j,k+1/2}^{n+(1-m)} \right) \quad (12)$$

Similar equations can be obtained for the other field components.

B. Absorbing Boundary Conditions

Using the anisotropic PML (APML) formulations of [5], (1) and (2) can be written in the APML region at the domain boundaries as

$$j\omega\varepsilon_0\varepsilon_r(\omega)\bar{\varepsilon}(\mathbf{r}, \omega)\mathbf{E}(\mathbf{r}, \omega) = \nabla \times \mathbf{H}(\mathbf{r}, \omega) \quad (13)$$

$$j\omega\mu_0\bar{\mu}(\mathbf{r}, \omega)\mathbf{H}(\mathbf{r}, \omega) = -\nabla \times \mathbf{E}(\mathbf{r}, \omega) \quad (14)$$

where $\bar{\varepsilon}(\mathbf{r}, \omega)$ and $\bar{\mu}(\mathbf{r}, \omega)$ are, respectively, the APML permittivity and permeability diagonal tensors defined as [5]

$$\bar{\varepsilon}(\mathbf{r}, \omega) = \bar{\mu}(\mathbf{r}, \omega) = \begin{bmatrix} S_y S_z & & \\ & S_x & \\ & & S_y \\ & & & S_x S_y \\ & & & & S_z \end{bmatrix} \quad (15)$$

with S_η ($\eta = x, y, \text{ or } z$) are given by

$$S_\eta = 1 + \frac{\sigma_\eta}{j\omega\varepsilon_0} \quad (16)$$

where σ_η is the APML conductivity profile along the η -coordinate designed to absorb the outgoing waves with minimal reflections [3]. To discretize (13) and (14), consider, as an example, the E_z -field component of (13):

$$j\omega\varepsilon_0\varepsilon_r(\omega) \frac{\left(1 + \frac{\sigma_y}{j\omega\varepsilon_0}\right)\left(1 + \frac{\sigma_x}{j\omega\varepsilon_0}\right)}{\left(1 + \frac{\sigma_z}{j\omega\varepsilon_0}\right)} E_z = \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} \quad (17)$$

equation (17) can be re-arranged as

$$j\omega\varepsilon_0 \left(1 + \frac{\sigma_x}{j\omega\varepsilon_0}\right) G_z = \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} \quad (18)$$

where G_z is given by

$$G_z = \frac{\left(1 + \frac{\sigma_y}{j\omega\varepsilon_0}\right)}{\left(1 + \frac{\sigma_z}{j\omega\varepsilon_0}\right)} D_z \quad (19)$$

and D_z is related to E_z through (6). Using the inverse Fourier transform relation, $j\omega \Rightarrow \partial/\partial t$, (18) and (19) can be written in the time domain as

$$\frac{\partial G_z}{\partial t} + \frac{\sigma_x}{\varepsilon_0} G_z = \frac{1}{\varepsilon_0} \left(\frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} \right) \quad (20)$$

and

$$\frac{\partial G_z}{\partial t} + \frac{\sigma_z}{\varepsilon_0} G_z = \frac{\partial D_z}{\partial t} + \frac{\sigma_y}{\varepsilon_0} D_z \quad (21)$$

Using the FDTD algorithm [2], (20) and (21) can be written in the discrete time domain as

$$G_{z,i,j,k+1/2}^{n+1} = \frac{\alpha_{x_i}^-}{\alpha_{x_i}^+} G_{z,i,j,k+1/2}^n + \frac{\Delta t}{\alpha_{x_i}^+ \Delta \varepsilon_0} \left[H_{y,i+1/2,j,k+1/2}^{n+1/2} - H_{y,i-1/2,j,k+1/2}^{n+1/2} - H_{x,i,j+1/2,k+1/2}^{n+1/2} + H_{x,i,j-1/2,k+1/2}^{n+1/2} \right] \quad (22)$$

$$D_{z,i,j,k+1/2}^{n+1} = \frac{\alpha_{z_j}^-}{\alpha_{z_j}^+} D_{z,i,j,k+1/2}^n + \frac{\alpha_z^{k+1/2}}{\alpha_{z_j}^+} \left[G_{z,i,j,k+1/2}^{n+1} - \frac{\alpha_{z_{k+1/2}}^-}{\alpha_{z_{k+1/2}}^+} G_{z,i,j,k+1/2}^n \right] \quad (23)$$

where $\alpha_{\eta_m}^\pm$, (for $\eta = x, y, \text{ or } z$), is given by

$$\alpha_{\eta_m}^\pm = 1 \pm \Delta t \sigma_{\eta_m} / 2\varepsilon_0 \quad (24)$$

After computing $G_{z,i,j,k+1/2}^{n+1}$ and $D_{z,i,j,k+1/2}^{n+1}$ from (22) and (23), respectively, $E_{z,i,j,k+1/2}^{n+1}$ can be obtained from (11) and (12). It is important to note that (22) and (23) can also be applied in the inner FDTD computational domain by setting the APML conductivity profiles ($\sigma_\eta(\eta = x, y, z)$) to zero. Similar expressions can be obtained for the other field components.

C. Parallelization Strategy

In the presented parallel algorithm, the computational domain is spatially partitioned into adjacent non-overlapping sub-domains using 2-D topology, in which the computational domain is divided into sub-domains along two directions. Fig.

1 shows a typical 2-D decomposition when 4 processors are used. To update the field components at the sub-domain boundaries, data from the neighboring sub-domains are needed. In this paper, the inter-processor communication among the neighboring processors is carried out by using the MPI library [7]. Fig. 2 shows the data need to be exchanged between neighboring sub-domains. For the communication purpose, ghost layers located at the edges of the sub-domains are used as shown in Fig. 2. It is important to note that as the APML field equations involve the same number of inter-processor communication operations as the conventional FDTD equations, as can be seen from (7), (22), and (23), the APML finite-difference equations can be used for the total computation domain by properly choosing the APML parameters. This makes the parallel FDTD algorithm easier to implement.

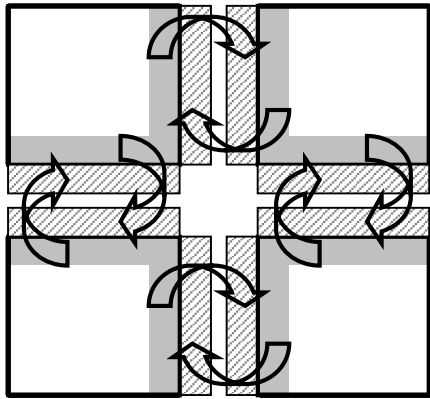


Fig.1: Computational domain partitioning using 2-D topology. Shaded and gray layers represent ghost layers and internal edges at the sub-domains, respectively.

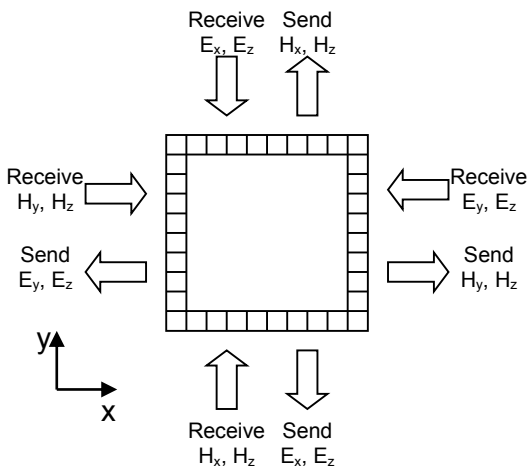


Fig. 2: Communications at the boundaries of a sub-domain for the 2-D topology.

Finally, the steps for the proposed parallel algorithm can be summarized as:

1. MPI initialization.
2. Reading of simulation parameters.
3. Creation of the 2-D topology.
4. At each time step perform the following:
 - 4.1 Exchange **H**-fields with the neighbor sub-domains by using the MPI library functions.
 - 4.2 Update the **E**-fields and other auxiliary variables in each sub-domain.
 - 4.3 Exchange **E**-fields with the neighbor sub-domains by using the MPI library functions.
 - 4.4 Update the **H**-fields in each sub-domain.
5. MPI finalization.

TABLE I: FDTD PARALLEL SYSTEM CHARACTERISTICS.

CPU	Pentium IV 2.20 GHz
Memory	512 Mbyte
Processor number	4, 8, or 16
Communication software	Message passing interface
Network interface	100Mbps Ethernet
Operating system	Windows XP
Compiler	C++

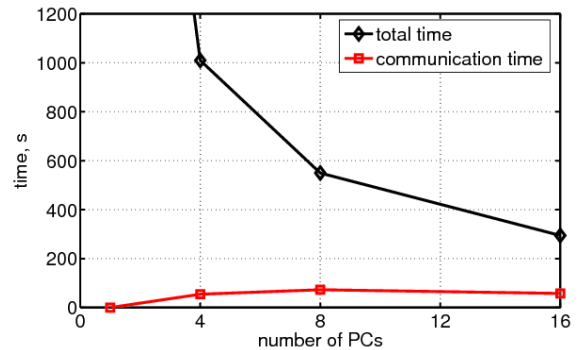


Fig. 3: Total simulation time and communication time of the proposed parallel algorithm

III. SIMULATION STUDY

The performance of the proposed parallel algorithm was studied for a three dimensional radiation problem. In this study, a *z*-polarized modulated Gaussian pulse with a carrier frequency of 20 PHz was excited at the center of $240\Delta \times 240\Delta \times 40\Delta$ computational domain, where $\Delta = \Delta x = \Delta y = \Delta z = 1 \times 10^{-10}m$. The computational domain was entirely composed of linear Lorentz material ($M = 2$) with a dielectric permittivity given by

$$\epsilon_r(\omega) = \epsilon_\infty + \frac{\Delta\epsilon\omega_0^2}{\omega_0^2 + j2\delta\omega - \omega^2} \quad (25)$$

where $\epsilon_\infty = \epsilon_r(\infty) = 1.0$, $\Delta\epsilon = \epsilon_s - \epsilon_\infty$, with $\epsilon_s = \epsilon_r(0) = 2.25$, $\omega_0 = 4 \times 10^{16}rad/s$ is the resonance radial frequency, and $\delta = 0.28 \times 10^{16}s^{-1}$ is the damping constant [17]. In this case, the coefficients of (3) are

$$a_0 = (\epsilon_\infty + \Delta\epsilon)\omega_0^2, a_1 = 2\epsilon_\infty\delta, a_2 = \epsilon_\infty,$$

$$b_0 = \omega_0^2, b_1 = 2\delta, \text{ and } b_2 = 1 \quad (26)$$

The computational domain was truncated by eight additional PML layers with a quadratic conductivity profile and with a theoretical reflection coefficient of 10^{-5} , as defined in [3]. The simulation time was carried out for the first 2500 time steps and the time step was taken as $\Delta_t = \Delta/(\sqrt{3}c/\sqrt{\epsilon_\infty})$, where c is the speed of light in vacuum. The parallel system used in this study was composed of 16 PCs interconnected through 100Mbps ethernet. Table I shows the characteristics of the proposed parallel system. Fig. 3 shows the total simulation time and the communication time of the proposed parallel algorithm.

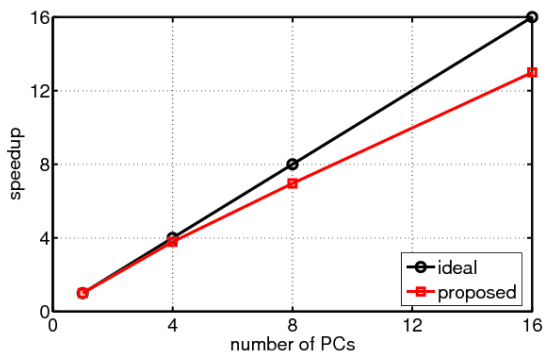


Fig. 4: Speed-up of the proposed parallel algorithm.

The performance of the proposed parallel algorithm was studied according to Speedup and efficiency factors. The speedup was calculated as

$$S(P) = T(1)/T(P) \quad (27)$$

where $T(1)$ is the time needed to solve the problem using one processor and $T(P)$ is the time needed to solve the same problem using P processors. The efficiency was calculated as

$$E(P) = S(P)/P \quad (28)$$

Figs. 4 and 5 show, respectively, the speedup and the efficiency of the proposed parallel algorithm. For the purpose of comparison, the ideal speedup and efficiency were also shown in Figs. 4 and 5. As can be seen from Fig. 4, almost linear speedup was obtained when the parallel code was run on less than four processors. Beyond this, the efficiency of the parallel system decreases. This is due to the fact that as the number of processors increases, the size of each sub-domain will be too small and hence the communication time becomes comparable to the computational time in the sub-domain. It is important to note that the performance of the parallel system can be improved further by using 3-D topology, which involves dividing the computational domain in the x , y , and z -directions [8].

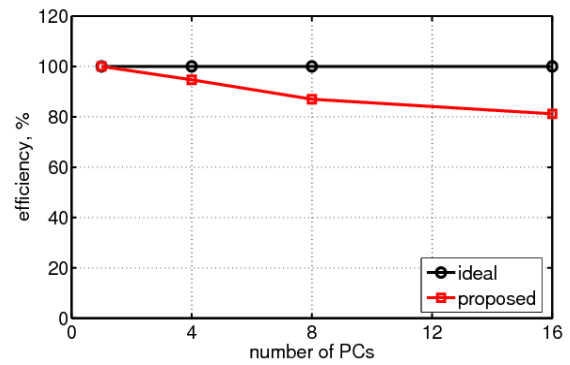


Fig. 5: Efficiency of the proposed parallel algorithm.

Finally, it should be noted that in the above example, the scalability of the proposed parallel algorithm was measured for a fixed problem size. Nevertheless, for some problems, parallel systems can also be used to solve larger problems. For such problems, the performance of the parallel algorithm is measured as the problem size scales proportionally to the number of processors. In this case, the computational problem size is kept constant per processor, while the number of processors increases. In the present study, the sub-domain size is kept fixed at $240 \times 240 \times 40$ per processor. Table II shows the scalability of the proposed parallel algorithm. As can be seen from these results, although the problem size is increased, there is a slight change in the total simulation time, which is due to the communication time between the processors. Hence, the problem size can be increased as the number of processors is increased.

TABLE II: SCALABILITY OF THE PROPOSED PARALLEL ALGORITHM FOR SCALED PROBLEM SIZE.

P	P_x	P_y	N_x	N_y	Computation time
1	1	1	240	240	3814.8
4	2	2	480	480	4083.0
8	4	2	960	480	4131.6
16	4	4	960	960	4275.6

IV. CONCLUSIONS

In this paper, full-wave parallel FDTD algorithm is presented for modeling electromagnetic wave propagation in dispersive open region problems. In the presented work, the problem geometry is divided into non-overlapping sub-domains using the 2-D topologies. It has been observed that the proposed parallel algorithm not only speed up computations but also increases the maximum solvable problem size. It is important to note that the presented formulations can be used for modeling electromagnetic waves interactions with human tissues like mobile phone radiations effect on human head. Finally, it should be noted that the simulations can be accelerated dramatically by using the graphical processing unit

(GPU) and employing the compute unified device architecture (CUDA) parallel programming model [18], and this issue is under investigations.

REFERENCES

- [1] K.S. Yee, "Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media", *IEEE Transaction on Antennas and Propagation*, vol. 14, no. 3, pp. 302-307, May 1966.
- [2] A. Taflove and S.C. Hangess, *Computational electrodynamics: The Finite-Difference Time-Domain Method*, 2nd ed., Norwood, MA: Artech-House, 2000.
- [3] J.P. Berenger, "A perfectly matched layer for the absorption of electromagnetic waves", *Journal of Computational Physics*, vol. 114, no. 2, pp. 185-200, Oct. 1994.
- [4] W.C. Chew, and W.H. Weedon, "A 3-D perfectly matched medium from modified Maxwell's equation with stretched coordinates," *Microwave and Optical Technology Letters*, vol. 7, no. 13, pp. 599-604, Sep. 1994.
- [5] S.D. Gedney, "An anisotropic perfectly matched layer absorbing medium for the truncation of FDTD lattices", *IEEE Transactions on Antennas and Propagation*, vol. 44, no. 12, pp. 1630-1639, Dec. 1996.
- [6] S. A. Cummer, "A simple, nearly perfectly matched layer for general electromagnetic media," *IEEE Microwave Wireless Component Letters*, vol. 13, no. 3, pp. 128-130, Mar. 2003.
- [7] W. Gropp, E. Lusk, and A. Skjellum, *Using MPI: Potable parallel Programming with the Message-Passing Interface*, 2nd ed., Cambridge, MA, MIT Press, 1999.
- [8] H. Hoteit, R. Sauleau, B. Philippe, P. Coquet, and J.P. Daniel, "Vector and parallel implementations for the FDTD analysis of millimeter wave planar antennas", *International Journal of High Speed Computing*, vol. 10, no. 2, pp. 209-234, Jun. 1999.
- [9] C. Guiffaut, and K. Mahdjoubi, "A parallel FDTD algorithm using the MPI library", *IEEE Antennas and Propagation Magazine*, vol. 43, no. 2, pp. 94-103, Apr. 2001.
- [10] D.-H. Sheen, K. Tuncay, C.-B. Baag, P. J. Ortoleva, "Parallel implementation of a velocity-stress staggered-grid finite-difference method for 2-D poroelastic wave propagation", *Computers and Geosciences*, vol. 32, no. 8, pp. 1182-1191, Oct. 2006.
- [11] O. Ramadan, "Three Dimensional MPI Parallel Implementation of the PML algorithm for truncating finite-difference time-domain grids," *Parallel Computing*, vol. 33, no. 2, pp. 109-115, Mar. 2007.
- [12] O. Ramadan and O. Akaydin, "Efficient parallel PML algorithms for truncating finite difference time domain simulations," *Electrical Engineering*, vol. 90, no. 3, pp. 175-180, Feb. 2008.
- [13] J.S. Ayubi-Moak, S.M. Goodnick, D. Stanzione, G. Speyer, and P. Sotirelis, "Improved parallel 3D FDTD simulator for photonic crystal", DoD HPCMP Users Group Conference, 14-17 July 2008, Seattle, WA, pp. 319-326.
- [14] J. Li, L.-X. Guo, H. Zeng, and X.-B. Han, "Message-passing-interface-based parallel FDTD investigation on the EM scattering from a 1-D rough sea surface using uniaxial perfectly matched layer absorbing boundary", *Journal of the Optical Society of America A*, vol. 26, no 6, pp. 1494-1502, Jun 2009.
- [15] O. Ramadan, "An Efficient MPI-Based Parallel Wave-Equation FDTD Algorithm for Dispersive Electromagnetic Applications," *The 5th International Conference on Information Technology, (ICIT'11)*, Amman, Jordan, May 11-3, 2011
- [16] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms and Applications*, 3rd ed., Englewood Cliffs, NJ: Prentice-Hall, 1995.
- [17] R. M. Joseph, S. C. Hagness, and A. Taflove, "Direct time integration of Maxwell's equations in linear dispersive media with absorption for scattering and propagation of femtosecond electromagnetic pulses," *Optics Letters*, vol. 16, no. 18, pp. 1412-1414, Sep. 1991.
- [18] Nvidia, CUDA C programming guide version 4.0, Nvidia Corporation, Santa Clara, CA, 2011.

The Dualism of Context in Ubiquitous Computing

Dennis Lupiana

Faculty of Computing, Information System and Mathematics
Institute of Finance Management
Dar Es Salaam, Tanzania

Fredrick Mtenzi

School of Computing
Dublin Institute of Technology
Dublin, Ireland

Brendan O'Shea

School of Computing
Dublin Institute of Technology
Dublin, Ireland

Abstract— Context-aware systems are fundamental for making the use of computing devices intuitive. These systems respond to their environments to facilitate seamless interactions between the users and their computing devices, and to make these devices less intrusive. Although it is more than a decade since context-aware systems were introduced, context is still not well understood within a context-awareness research community. Although there are numerous definitions of context, these definitions refer to context as an input or as a derivable. The majority of researchers believe any input that makes a context-aware system to accomplish its task is a context. In the contrast, there is handful of researchers who believe context is derived from more than one inputs. This paper aims to provide a clear meaning of context and consequently to resolve the differences between researchers regarding context. In particular, this paper answers the most fundamental, but yet the most avoided, question; *what is context?*

Keywords— *context, situation, context-awareness system; ubiquitous computing; context-aware architecture*

I. INTRODUCTION

The Ubiquitous Computing (UbiComp) paradigm has inspired the invention of numerous computing devices. Although these devices offer the users many convenient ways of accomplishing their everyday tasks, it remains a challenge for the users to use them effectively. This is more challenging as these devices are mobile. The users' working environments become open and hence less predictable. The users and their devices enter and leave different working environments where different settings and computing needs may be required. This makes interacting with devices difficult and more time consuming.

In response to these challenges, a Context-Awareness research strand emerged. The main focus of this strand is to investigate different principles, methodologies and techniques required to develop software systems that can *adapt* to their dynamic environments and the users' computing needs [1]. These systems are called *context-aware systems*. Initial context-aware systems used location or identity information to automatically provide users' computing needs. To date there

are many context-aware systems, each exploiting different aspects of the real world.

Central to a context-aware system is *context*. Despite of its importance, context is still not well understood within the context-awareness research community. As a result, context means differently to different researchers. Although there are numerous definitions of context, there are two notable interpretations of context; *context as an input* and *context as a derivable*. The majority of researchers believe any input that makes a context-aware system to accomplish its task is context. In the contrast, there is handful of researchers who believe context is derived from more than one inputs.

This paper aims to provide a clear meaning of context and consequently to resolve the differences between researchers regarding context. In particular, this paper answers the most fundamental, but yet the most avoided, question; *what is context?* This is not the first time this question is raised. Dey and Abowd [2] and Zimmermann and his colleagues [3] have raised and attempted to address this question. The majority of researchers avoid defining context and instead adopts the

existing definitions of context. As contended by [4], defining context is difficult and hence researchers prefer to adopt the existing definitions. Others argue that the definition of context is not important but how context is used is. Since context is central to context-aware systems, in this paper we argue that a clear understanding of context is important.

The rest of this paper is organised as follows. The background of context-awareness computing is provided in section II where the reason for the different interpretations of context is outlined. The explanation of the two notable interpretations of context is provided in section III and IV. The discussion of these interpretations is provided in section V. Section VI discusses the implications and consequences of these two interpretations. Section VII provides a conclusion of this paper and the future work.

II. BACKGROUND ON CONTEXT-AWARENESS

The pioneers of Context-Awareness computing [1] define context-aware system as a computing system that examines and reacts to individual's context. To examine is to scrutinise or analyse while to react is simply to respond to something. Hence, according to Schilit and his colleagues, context-aware systems are computing systems that analyse someone's context before responding to it. This implies that context is dynamic and thus context-aware systems should be adaptive to these dynamics. As Schilit and his colleagues assert, the constantly changing execution environment is a significant aspect of context-awareness. This leaves us with many questions but the most important one is; *what is context?*

Unfortunately, Schilit and his colleagues provide us with no definition of context. Instead, they assert that where you are (location), who are you with (other people) and what resources are nearby (accessible devices) are important aspects of context. As Schilit and his colleagues argue, the little information covering someone's proximate environment is the most important in context-awareness. Clearly, these aspects are "ingredients" and context is an end product. Hence, context-aware systems should use these "ingredients" as inputs to determine and subsequently to respond to someone's context. This implies that context-aware systems should be responsive to context and not to individual inputs.

In the contrary, initial context-aware systems [5-9] are described to be responsive to implicit inputs such as identity and location. Weiser [10], for instance, describes a system that opens a door to the right badge wearer. In these systems, responses are predetermined and hence inputs are used as *cues*. Although Schilit and his colleagues [11] later call for a broader view of context, the description of the initial context-aware systems had already caused a considerable divide among researchers. While the majority of researchers believe that an input to a context-aware system is a context, few argue that context is derived from more than one input.

III. CONTEXT AS AN INPUT

Dey and Abowd [2] define a context-aware system as a computing system capable of providing information and/or services relevant to the user's task. This definition raises a

question as to how would a context-aware system know what task a user is involved in? According to [2], context-aware systems do not know users' tasks but are programmed to provide relevant information and/or services in a task. To automate the latter process, these systems are developed to respond to cues. In a context-aware tour guide system, for example, a tourist is provided with relevant information about a site when approaching the site. In this example, the location of the site is a cue.

Dey and Abowd [2] refer to cues as context and define it as *any* information that can be used to characterise a situation of an entity, where an entity can be a user or any other object. This definition implies that context can be one or a set of inputs. In the tour guide example, for instance, location is a context because it is used to characterise the situation of the user. This is a fairly reasonable definition of context since it is open-ended. It provides flexibility to system developers to enumerate contexts as required by their systems. According to their example, a situation is a task that a user needs to accomplish. Hence, any information that is necessary for the system to accomplish this task is context. Although this definition is broad, as argued by [3], it addresses the criticism that it is impossible to enumerate exhaustive list of context.

Although it is more than a decade since this definition was proposed, it is still used in the recent work. Soylu and his colleagues [12], Liu [13] and van de Westelaken and his colleagues [14], for instance, used this definition. As stated by [4], defining context is difficult and hence the majority of researchers prefer to adopt existing definitions. Although few researchers have attempted to define context, in many cases these definitions are variations of the definition provided by [2]. Chen and Kotz [15], for instance, define context as a set of environmental states and settings. Similarly, Chen [16] defines context by replacing 'any information' from the definition provided by [2] with a list of attributes and entities within a physical environment. Zimmermann and his colleagues [3] attempt to narrow the definition provided by [2].

In this interpretation, context is regarded as an input to a context-aware system. In many cases, context is referred to an attribute of an entity, which is essential for a context-aware system to accomplish a specific task. Consequently, as shown in figure 1, context-aware system is regarded as a system that monitors (P_1) its environment and responds (P_2) to inputs from its sensors.

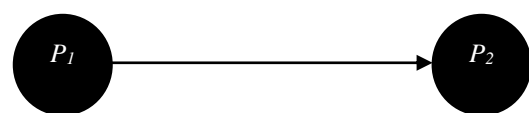


Fig. 1. Key processes of context-aware systems in this category

Two concepts from the description of context-aware systems from [1] and [2], however, are not blending well into this interpretation of context. First is the analogy of human perception and information within someone’s proximate environment. Human being uses information around them to understand their environments and hence to respond appropriately. Context-aware systems in this category, however, respond to individual pieces of information. Hence it is unclear on how context-aware systems use information from their surroundings to respond appropriately. Second is the idea of dynamism of context and how it is used in context-aware systems of this category.

IV. CONTEXT AS A DERIVABLE

“In this model computation does not occur at a single location and in a single context, as in desktop computing, but rather spans a multitude of situations and locations ...” [1].

It is evident from the excerpt that context is dynamic and it occurs in a location. This excerpt implies that location is not a context but it is part of context. As noted by [1], the dynamic environment of devices is the driving factor for designing context-aware systems.

Respond and *adapt* are two different terms and hence they should be carefully used, especially in context-awareness. In English, to respond means to act on return while to adapt means to modify or adjust to new conditions. Thus, in a responsive system the relation between inputs and outcomes is binary; the outcomes depend on whether the inputs exist or not. To be adaptive, however, a system should have a certain degree of correctness. This implies that an adaptive system should be able to examine or analyse its inputs. As noted by [2], “adapting to context” means a context-aware system can modify its behaviours accordingly.

Chen and Kotz [15] and Kaenampornpan [17] argue that a context-aware system should combine various inputs. Kofod-Petersen [18] also argues that if a context-aware system is unable to *reason* about various inputs, it cannot be adaptive to its context. Likewise, [12] argue that a context-aware system should exhibit intelligence. This implies that context is dynamic and hence context-aware systems should be able to contemplate different aspects of their environments before responding. This view of context-aware systems correlates with the definition of context-aware systems provided by [1]. These systems do not respond to individual inputs but after analysing these inputs. Hence context in these systems is derived from more than one input. Consequently, as shown in figure 2, a context-aware system is regarded as a system that monitors (P_1) its environment, analyses (P_2) its inputs and responds (P_3) to ongoing context. Hence a change of an input implies a change to an ongoing context. Thus, the dynamism of context is subject to different aspects of an environment.

Fig. 2. Key processes of context-aware systems in the ‘context as derivable’ category.

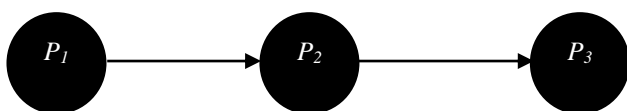
This also explains why human perception and information within someone’s proximate environment is relevant in context-awareness. Human perception can be summarized as a three-phase process, involving sensing of surroundings, interpreting of stimulus from sensory organs, and inferring what is going on. To know what is going on, human beings use their past experience about a phenomenon. In context-awareness, a similar argument has been raised. Bolchini and his colleagues [19], for instance, argue that context-awareness involves applying past experience to the available facts within the environment to understand what is happening. Similarly, [17] argues that a context-aware system should possess prior knowledge about situations. Therefore, like human beings, context-aware systems should also apply previous knowledge about contexts within their environments.

V. DISCUSSION

As explained in section III and IV, the term context has two notable interpretations in the context-awareness research community. On one camp context is referred to as any input to a context-aware system while on the other camp context is referred to as a product of a context-aware system after combining more than one input. One camp argues that any information is a context as long as it affects how a context-aware system operates. Hence, depending on a context-aware system, context can be one or more pieces of information. In contrast, the other camp argues that context is derived after analysing more than one input. In this camp context is a collective term used to describe circumstances of a user in the real world environment.

Let us assume that we have two context-aware systems; one that automatically opens a door to a right badge wearer and the other that remotely switches ON a user’s computer when a user enters a room. Both of these systems depend on one input, which is the identity of the users’ badges, but react differently. The user’s goal in the first system is to open a door while the user’s goal of the other system is to switch ON her/his computer when enters a room. Hence each of these systems responds to a user’s goal. ID of the badge is used by these systems as a *cue*; for automating the process of opening the door or switching ON a computer. Thus, a context-aware system is developed to respond to user’s goals and not to the inputs of the system.

Since context-aware systems respond to users’ goals, these goals can be referred to as context. This type of context, however, is static as it is predefined by developers of context-aware systems. Consequently, as [2] found a decade ago, the majority of the existing context-aware systems are *responsive* rather than *adaptive* to their environments. Recently, we also



arrived to similar findings [23]. The difference between these is that currently context-aware systems respond to more than one input. Hence, if a context-aware system is developed to automatically open a door to a right badge wearer, this system will open the door even when the badge wearer is not intending to enter the room. Hence, like the initial context-aware systems, these systems use inputs as cues.

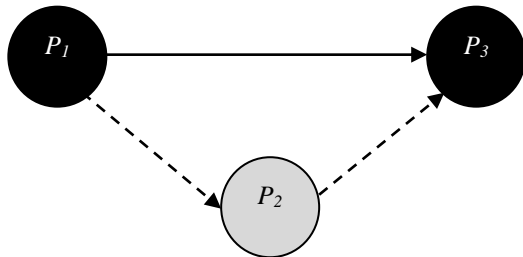


Fig. 3. Key processes of context-aware systems.

From the definition by [1], context-aware system should be capable of *monitoring* (P_1), *analysing* (P_2) and *responding* (P_3) to context as shown in figure 3. As asserted by [1], the constantly changing execution environment is a significant aspect of context-awareness. Hence all these processes play important roles but P_2 is the most important one as it determines context of a user. Thus, if at time T_i a system responded to one context then through P_2 the system would be able to respond to a different context at time T_{i+n} . Pieces of information gathered by the system, through P_1 , is used as an input to analyse the individual's context. The majority of the existing context-aware systems, however, are developed with predefined context and subsequently without P_2 . As a result, these systems cannot adapt to changes that occurs within their environments.

VI. IMPLICATIONS AND CONSEQUENCES

A. Inputs to Context-Aware Systems are Context Parameters

It is evident from the discussion that inputs to context-aware systems are parameters to these systems. Hence we argue that any information required by a context-aware system is a context parameter. We define context parameter as a piece of meaningful information that has an impact on a context-aware system. This information may be interpreted from data captured by a sensor or acquired directly from other sources such as a network or application software. The name of the owner of a device interpreted from the device's ID captured by a sensor, for instance, is a context parameter.

A widely used synonym of a context parameter is contextual information. This term, however, is interchangeably used with singular and plural meaning. Gu and colleagues [24]

and Chen [16], for instance, refer to identity, location or time as contextual information while [25] and [26] refer to a set of context parameters as contextual information. Hence, to avoid this confusion, we prefer to use the term context parameter.

B. Context is What a System is Programmed to Do

It is evident from the discussion that context-aware systems respond by accomplishing whatever are programmed to do. Inputs are used by context-aware systems as cues to automate whatever task a context-aware system is programmed to do. To avoid contradictions with context, as it is widely used, [15] refer to this kind of context as a *high level* context. Gellersen and colleagues [20] refer to this kind of context as a *situational context*. Barkhuus [21] refers to this kind of context as a human context. Similarly, [22] refers to this kind of context as *activity* or *situation*. Likewise, we [23] refer to this kind of context as *situation*.

This paper adopts the definition of a situation from [18] who defines a situation as a social setting, such as a meeting, where the users involved want to achieve various goals. This definition of situation differs from that of [2], [26] and [27] as is not confined to a particular task. This definition emphasises meaningful interactions between relevant entities required to sufficiently describe the real world environment that is of interest to the users and their devices. A situation provides a detailed picture of the real world environment whereas a context parameter provides an aspect of the real world environment.

C. Architectural Support are Key to Context-Awareness

As [2] and [23] found, currently the majority of the existing context-aware systems are responsive rather than adaptive to their changing environments. These systems respond appropriately only to contexts that are programmed with. In most cases, any change within a physical environment does not affect how these systems should respond. For instance, a context-aware system that automatically displays presentation slide will continue to display a slide with sensitive information even if an unauthorized person enters a room. Hence, should developers change how context-aware systems are currently implemented to accommodate the adaptive nature of context-aware systems?

For more than two decades context-aware systems have been implemented for a specific purpose. As the pioneers of context-aware computing have outlined, these systems provide and/or gather some sort of information or services or trigger some actions on their host devices. Therefore, changing how these systems should be implemented runs counter to the very fundamental principle of their existence. Developers can attempt to implement general purpose context-aware systems but they should be aware that there is no a killer-application.

The best that researchers can do is to develop knowledge-driven context-aware architectures. Instead of context-aware architectures to be *passive*, as the majority, these architectures should be *active*. In addition to acquiring and translating data from sensors and other sources, these architectures should be able to use numerous context parameters to determine ongoing

context. Hence, instead of context-aware architectures to be sharing context parameters with context-aware systems, these architectures should be sharing knowledge of what is happening. This will significantly increase application of context-awareness because even the most miniature and resource-limited devices would be made aware of their ongoing contexts.

To walk the walk, we have developed one of such architectures in the School of Computing at the Dublin Institute of Technology. In addition to monitoring and interpreting data gathered by sensors, our architecture called Knowledge-driven Distributed Architecture (KoDA) reasons about available information to recognise ongoing context. With KoDA, the implementation of context-aware systems is significantly simplified as they can be implemented without inference rules. This frees developers from the hurdles of learning knowledge representation languages required to represent inference rules. It also removes redundancy of inference rules as in KoDA these rules are centralised.

VII. CONCLUSION

In this paper we have reviewed relevant literature in context-awareness computing with the aim of providing a clear meaning of context and consequently to resolve the differences between researchers. In particular, this paper answers the most fundamental, but yet the most avoided, question; *What is context?*. We categorised the existing definitions of context into the two notable interpretations of context; context as an input and context as a derivable. The 'context as an input' interpretation of context qualifies individual pieces of information that have effect on how a context-aware system operates to context. In contrast, the 'context as a derivable' interpretation of context refers to context as to what that is derived after a context-aware system processes its inputs.

This paper is in favour of the 'context as a derivable' interpretation of context. This interpretation implies that context is what a context-aware system is programmed to accomplish. This interpretation also implies that context, as is used by the majority of the researchers, is a context parameter. It is an input that is essential for a context-aware system to accomplish its tasks. This *programmed* context is static and hence runs counter to the reasons for inventing context-aware systems in the first place. It is possible to develop context-aware systems that utilise data from numerous sensors and other sources to determine and subsequently to adapt to their environments. This requires huge processing power, which not all hosting computing devices have. Additionally, the form factor of the majority of hosting devices is not flexible to accommodate new sensors as are invented. Hence, we argue that architectural solutions are required in order to accommodate the adaptive nature of context-aware systems.

We have briefly described our Knowledge-driven Distributed Architecture (KoDA) which despite of monitoring and interpreting information from sensors, it reasons about the available information to recognise ongoing context. Among the benefits of KoDA includes the simplification of the process of implementing context-aware systems and the removal of

redundancy of inference rules. In KoDA, inference rules are not represented in context-aware systems. This significantly simplified the process of implementing context-aware systems as developers are freed from the hurdles of learning knowledge representation languages required to represent inference rules. In KoDA, inference rules are centralised and hence removes redundancy of inference rules.

CONCLUSION

- [1] Schilit, B., Adams, N. & Want, R. (1994). Context-aware computing applications. In *Mobile Computing Systems and Applications, 1994. WMCSA 1994. First Workshop on*, 85 - 90.
- [2] Dey, A. & Abowd, G. (2000a). Towards a better understanding of context and context-awareness. In *CHI 2000 workshop on the what, who, where, when, and how of context-awareness*, 304 - 307.
- [3] Zimmermann, A., Lorenz, A. & Oppermann, R. (2007). An operational definition of context. In *Modeling and using context*, 558 - 571, Springer.
- [4] Baldauf, M., Dustdar, S. & Rosenberg, F. (2007). A survey on context-aware systems. *International Journal of Ad Hoc and Ubiquitous Computing*, 2, 263 - 277.
- [5] Want, R., Hopper, A., Falcao, V. & Gibbons, J. (1992). The active badge location system. *ACM Trans. Inf. Syst.*, 10, 91 - 102.
- [6] Schilit, B. & Theimer, M. (1994). Disseminating active map information to mobile hosts. *Network, IEEE*, 8, 22 - 32.
- [7] Pascoe, J. (1998). Adding generic contextual capabilities to wearable computers. In *Wearable Computers, 1998. Digest of Papers. Second International Symposium on*, 92 - 99, IEEE.
- [8] Ryan, N.S., Pascoe, J. & Morse, D.R. (1998). Enhanced reality fieldwork: the context-aware archaeological assistant. In *Computer applications in archaeology*, Tempus Reparatum.
- [9] Abowd, G., Atkeson, C., Hong, J., Long, S., Kooper, R. & Pinkerton, M. (1997). Cyberguide: A mobile contextaware tour guide. *Wireless Networks*, 3, 421 - 433.
- [10] Weiser, M. (1991). The computer for the 21st century. *Scientific American*.
- [11] Schilit, B.N., LaMarca, A., Borriello, G., Griswold, W.G., McDonald, D., Lazowska, E., Balachandran, A., Hong, J. & Iversen, V. (2003). Challenge: Ubiquitous location-aware computing and the place lab initiative. In *Proceedings of the 1st ACM international workshop on Wireless mobile applications and services on WLAN hotspots*, 29 - 35, ACM.
- [12] Soylu, A., Causmaecker, P.D. & Desmet, P. (2009). Context and adaptivity in pervasive computing environments: Links with software engineering and ontological engineering. *Journal of Software*, 4, 992 - 1013.
- [13] Liu, H. (2010). Biosignal controlled recommendation in entertainment systems. *Technische Universiteit Eindhoven, Eindhoven*, 1 - 133.
- [14] van de Westelaken, R., Hu, J., Liu, H. & Rauterberg, M. (2011). Embedding gesture recognition into airplane seats for in-flight entertainment. *Journal of Ambient Intelligence and Humanized Computing*, 2, 103 - 112.
- [15] Chen, G. & Kotz, D. (2000). A survey of context-aware mobile computing research. Tech. rep., Citeseer.
- [16] Chen, H. (2004). *An Intelligent Broker Architecture for Pervasive Context-Aware Systems*. Ph.D. thesis, University of Maryland.
- [17] Kaenampornpan, M. (2009). *A Context Model, Design Tool and Architecture for Context-Aware Systems Design*. Ph.D. thesis, Department of Computer Science, University of Bath.
- [18] Kofod-Petersen, A. (2007). *A Case-Based Approach to Realising Ambient Intelligence among Agents*. Ph.D. thesis, Department of computer and Information Science, Norwegian University of Science and Technology.

- [19] Bolchini, C., Curino, C.A., Quintarelli, E., Schreiber, F.A. & Tanca, L. (2007). A data-oriented survey of context models. *SIGMOD Rec.*, 36, 19 - 26.
- [20] Gellersen, H.W., Schmidt, A. & Beigl, M. (2002). Multi-sensor contextawareness in mobile devices and smart artifacts. *Mobile Networks and Applications*, 7, 341 - 351.
- [21] Barkhuus, L. (2005). *The Context Gap: An Essential Challenge to Context-Aware Computing*. Ph.D. thesis, The IT University of Copenhagen.
- [22] McKeever, S. (2011). *Recognising Situations Using Extended Dempster-Shafer Theory*. Ph.D. thesis, School of Computer Science and Informatics, National University of Ireland.
- [23] Lupiana, D. (2015). *A Knowledge-driven Distributed Architecture for Context-Aware Systems*. Ph.D. thesis, School of Computing, Dublin Institute of Technology.
- [24] Gu, T., Pung, H.K. & Zhang, D.Q. (2005). A service-oriented middleware for building context-aware services. *Journal of Network and computer applications*, 28, 1 - 18.
- [25] Ye, J., Coyle, L., Dobson, S. & Nixon, P. (2007). Using situation lattices to model and reason about context. In *Modeling and Reasoning in Context (MRC) with Special Session on the Role of Contextualization in Human Tasks (CHUT) which is held in conjunction with CONTEXT*, 1 - 12.
- [26] Henricksen, K. (2003). *A Framework for Context-Aware Pervasive Computing Applications*. Ph.D. thesis, School of Information Technology and Electrical Engineering, The University of Queensland.
- [27] Ranganathan, A. & Campbell, R.H. (2003a). *An infrastructure for context-awareness based on first order logic*. *Personal and Ubiquitous Computing*, 7, 353 - 364.

Anatomy of the Parallel Tree Based Strategy for High Strength Interaction Testing

Mohammad F. J. Klaib
Computer Science Department
College of Computer Sciences and Engineering
Taibah University
Madina, Kingdom of Saudi Arabia
Email: mklaib@taibahu.edu.sa
mom_klaib@yahoo.com

Abstract—Software and hardware testers concentrate on how to minimize the time involved in testing at the same time to ensure that the system is also tested well and made acceptable. This paper has enhanced and explained in details our previous strategy “A Tree Based Strategy for Test Data Generation and Cost Calculation for Pairwise Combinatorial Interaction Testing” to work effectively in parallel and to go beyond pairwise testing. The proposed strategy can now support a parallel 2-way and general multi-way combinatorial interaction test data generation based on two algorithms; a parallel tree generation algorithm which generates the test cases and a parallel T-way cost calculation algorithm which is used in constructing test suites with minimum number of test cases. Both strategies have been explained here in details.

Keywords— *Parallel algorithms, Software testing, Hardware testing, Multi-way testing*

I. INTRODUCTION

A well-tested product or service is necessity to ensure customer's satisfaction. However, exhaustive testing is unaffordable due to combinatorial explosion problem. Combinatorial explosion in testing may occur for configurable systems. When systems under test have many configuration parameters, each with several possible values, testing each configuration is sometimes infeasible.

Combinatorial interaction testing has been one of the methods used to minimize the size and the time involved in testing [28-32], at the same time to ensure that the system is also tested well and made acceptable. The combinatorial interaction testing approach can reduce the number of test cases by systematically selecting a subset from an exhaustive testing combination based on the strength of parameter interaction [9-14, 23,27]. Basic combinatorial interaction testing which is called pairwise or 2-way testing [4-8] provides a systematic approach to identify and isolate faults since many faults are caused by unexpected 2-way interactions among system factors. Empirical investigations have concluded that from 50 to 97 percent of software faults [1, 6, 9, 15, 24- 26] could be identified by pairwise combinatorial interaction testing. However, what about the remaining faults? Especially, in case of highly interactive systems which have a number of interactions with higher strength. How many failures could be triggered only by an unusual interaction involving more than two parameters? Investigations have found that many faults were caused by a single parameter, a smaller proportion resulted from an interaction between two parameter values, and progressively fewer were triggered by 3-6 way interactions [3,18- 22].

Therefore, to ensure a high quality testing of complex applications, it is necessary to generate test suites for higher degree T-way interactions. T-way testing [3, 18, 19, 20, 21, 22] requires every combination of any T parameter values to be covered by at least one test, where T is referred to as the strength of coverage. If all faults in a system can be triggered by a combination of T or fewer parameters, then testing all T-way combinations of parameters can provide high confidence that nearly all faults have been discovered. A number of studies have shown combinatorial methods to be highly effective for software and hardware testing.

Large and/or computationally expensive optimization problems sometimes require parallel or high-performance computing systems. Parallel algorithms have been applied to problems such as weather and climate modelling, bioinformatics analysis, logistics and transportation, and engineering design. Furthermore, commercial applications are driving development of effective parallel software [16, 17, 22, 26] for large-scale applications such as data mining and computational medicine. In the simplest sense, parallel computing involves the simultaneous use of multiple computer resources to solve a computational problem. In this paper we have enhanced our previous strategy “A Tree Based Strategy for Test Data Generation and Cost Calculation” [24, 25, 26] to work in parallel and to go beyond pairwise (2-way) testing. The proposed strategy can now support a parallel and general T-way combinatorial test data generation involving uniform and non uniform parametric values. The proposed strategy is based on two algorithms; a parallel tree based test data generation algorithm which generates all the test cases, and a parallel T-way cost calculation algorithm which is applied to construct T-way test suites with minimum number of test cases.

The remainder of this paper is organized as follows. Section 2 explains the parallel tree generation and the proposed iterative T-way cost calculation strategy with an example. Section 3 gives parallel tree generation algorithms for test case generation and explains its advantages. Section 4 presents the parallel, iterative, T-way cost calculation algorithm for T-way test suites generation, with its working explained. Finally, Section 5 gives the conclusion.

II. THE PROPOSED STRATEGY

The proposed strategy constructs the tree based on the parameters and values given. It constructs every branch of the tree in parallel. The number of branches the tree has depends on the number of values of the first parameter i.e. if the first parameter has 3 values then the tree also would have 3 branches. Therefore every branch construction starts by getting one value of the first parameter i.e. branch T1 gets the first value, T2 gets the second value and so on. After the base branches are constructed one child thread is assigned to every branch and the further construction takes place in a parallel manner. Each of the branches considers all values of all the other parameters two, three,N where N is the total number of parameters. All the branches consider the values of the parameters in the same order. The following simple system with parameters and values, illustrates the concept as shown below:

- Parameter A has two values A1 and A2
- Parameter B has one value B1
- Parameter C has three values C1, C2 and C3
- Parameter D has two values D1 and D2

We have given the illustration for minimum test suite construction of 2-way and 3-way combinatorial interactions testing using our algorithm, for the system mentioned. The algorithm starts constructing the test-tree by considering the first parameter. As the first parameter has two values the tree is said to have two main branches with the first branch using A1 and the second branch using A2. Then each of the branches is

constructed in parallel by considering all the values of the second parameter, then the third and fourth and so on. When the branches are fully constructed the leaf nodes gives all the test cases that has to be considered for cost calculation. Since all of the branches are constructed in parallel there is a significant reduction in time. Fig. 1 shows the test tree for the system below.

Fig. 1 above shows how the test-tree would be constructed. The test cases generated by the first branch are stored in the lists T₁ and the test cases generated by the second branch are stored in T₂ respectively. i.e. (A1,B1,C1,D1), (A1,B1,C1,D2), (A1,B1,C2,D1), (A1,B1,C2,D2), (A1,B1,C3,D1), (A1,B1,C3,D2) are stored in T₁, and (A2,B1,C1,D1), (A2,B1,C1,D2), (A2,B1,C2,D1), (A2,B1,C2,D2), (A2,B1,C3,D1) and (A2,B1,C3,D2) are stored in T₂.

Once the parallel tree construction is over we are ready with all the test cases to start the parallel iterative cost calculation. In this strategy the cost of the leaf nodes in each of the lists are calculated in parallel in order to reduce the execution time. The cost of a particular test case is the maximum number of T-way combinations that it can cover from the covering array. At First, the algorithm starts by constructing the covering array, for all possible T-way combinations of input variables, if T equals 2 i.e. [A & B], [A & C], [A & D], [B & C], [B & D] and [C & D]. The covering array for the above example has 23 pairwise interactions as shown in Table 1, which has to be covered by any test suite generated, to enable a complete pairwise interaction testing of the system.

Once the covering array is generated the algorithm starts to include all tree branches. which might definitely give the maximum Wmax cost into the test suite. Then these test cases are deleted from the tree branches lists T₁ and T₂, and the corresponding pairs covered by it in the covering array are also deleted. In the third step, the main thread in the algorithm invokes a number of child threads equal to the number of values of the first parameter and calculates the cost of all the test cases

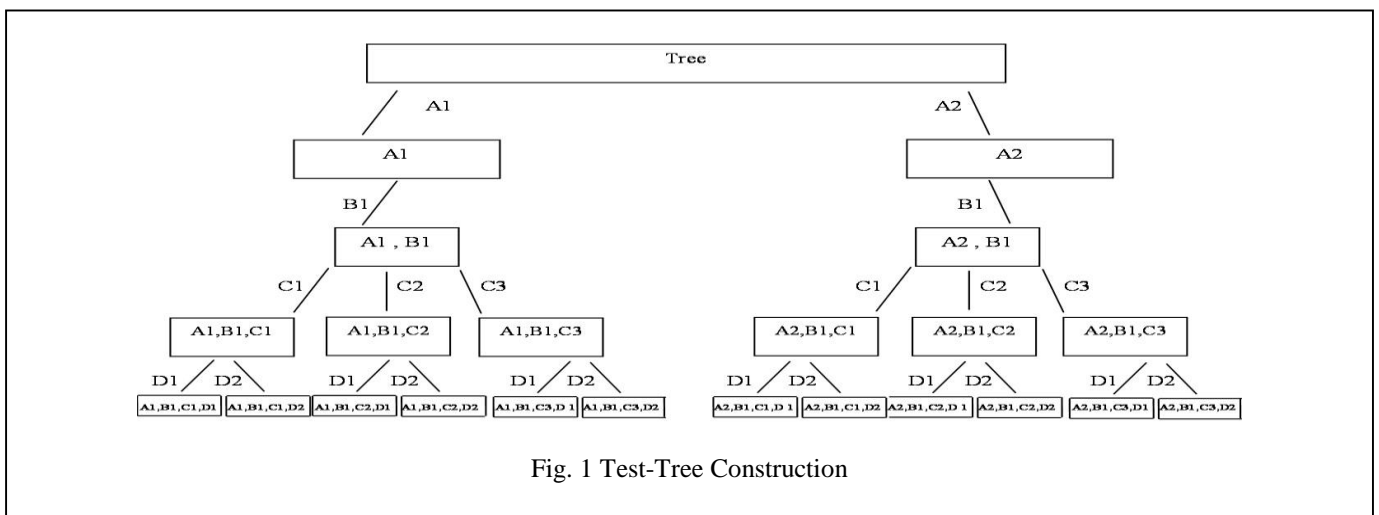


Fig. 1 Test-Tree Construction

in each of the branches in a parallel fashion. Each child thread stores all the test cases with the Wmax value from its corresponding branch into a separate sub-list. The child thread that finishes calculating the cost of all the test cases in its branch first locks the covering array. This thread then looks into its sub-list and includes the test cases stored in it into the test suite only after confirming that the test case definitely has the maximum cost or Wmax value. Then the test cases included in the test suite are deleted from the tree branches list and sub-list, and the corresponding pairs that these cover are deleted from the covering array.

A1, B1,C3	A2, B1,D1	A1, C2, D1	B1,C2, D1
A2, B1,C1	A2, B1,D2	A1, C2, D2	B1,C2, D2
A2, B1,C2		A1, C3, D1	B1,C3, D1
A2, B1,C3		A1, C3, D2	B1,C3, D2
		A2, C1, D1	
		A2, C1, D2	
		A2, C2, D1	
		A2, C2, D2	
		A2, C3, D1	
		A2, C3, D2	

TABLE 1. PAIRWISE COVERING ARRAY.

A with B	A with C	A with D	B with C	B with D	C with D
A1,B1	A1,C1	A1, D1	B1,C1	B1, D1	C1, D1
A2,B1	A1,C2	A1, D2	B1,C2	B1, D2	C1, D2
	A1,C3	A2, D1	B1,C3		C2, D1
	A2,C1	A2, D2			C2, D2
	A2,C2				C3, D1
	A2,C3				C3, D2

The other threads wait in a queue until the execution of the first thread is over, after which these threads resume their execution in the order in which they are queued. These threads on resumption re-evaluate the test cases in their sub-list to confirm that these test cases have the Wmax value before including these into the test suite. Thus in the first iteration all the test cases with the maximum Wmax value from all the branches are included in the test suite. Now the Wmax value is decremented by one and the same parallel execution of all the threads continue until all the pairs in the covering array are covered. For the above example all the test cases which are included in the test suite are identified in four iterations and there are six such test cases. Table 2 shows how the cost calculation works iteratively to generate the test suite. The same test suite gets generated if a sequential execution of the above algorithm takes place.

As the pairwise test suite is generated, we can generate the test suite for 3-way combinatorial interactions and so on the forth until (n-1) way combinatorial interaction test suites are generated. To illustrate the 3-way test suite generation, again the whole process starts by constructing the 3-way covering array and the iterative, parallel cost calculation of the test cases in the various branches as explained before. Table 3 shows the covering array for 3-way combination i.e. [A, B, C], [A, B, D], [A, C, D] and [B, C, D], for the example in Fig. 1. The covering array for the above example has 28 3-way interactions which have to be covered by any test suite generated, to enable a complete 3-way interaction testing of the system. Table 4 shows how the cost calculation works iteratively to generate the test suite. Table 4 also shows the order in which the various test cases are actually included in the test suite.

TABLE 3 3-WAY INTERACTION COVERING ARRAY.

A, B, C	A, B, D	A, C, D	B, C, D
A1, B1, C1	A1, B1, D1	A1, C1, D1	B1,C1, D1
A1, B1,C2	A1, B1,D2	A1, C1, D2	B1,C1, D2

III. PARALLEL TREE GENERATION ALGORITHMS FOR TEST CASE GENERATION

A. Tree Generation Algorithm for Main Thread

Input: A set of parameters and the values of the corresponding parameters

Output: Lists of test cases. Each list holds the Fig. 4 Cost Calculation Algorithm

Test cases generated by the tree in one particular branch of that tree.

Begin

X = number of values of first parameter p1

{For the first parameter p1}

T_i=V_i, where i=1,2,3,...,X/ parameter p1 has X values

If N=1 then stop and exit;

Create X threads with unique thread ids. Assign each T_i to a separate child thread and execute all the child threads in parallel

Wait for the termination of all the threads to get the results from all the branches.

End

TABLE 2. GENERATED TEST SUITE FOR PAIRWISE COMBINATORIAL INTERACTION.

Test Case No.	Test Case	Iteration/Child Thread No.	Max Weight	Covered pairs
T1	A1,B1,C1,D1	1/1	6	[A1,B1][A1,C1][A1,D1] [B1,C1][B1,D1][C1,D1]
T10	A2,B1,C2,D2	½	6	[A2,B1][A2,C2][A2,D2] [B1,C2][B1,D2][C2,D2]
T6	A1,B1,C3,D2	2/1	4	[A1,C3][A1,D2] [B1,C3][C3,D2]
T11	A2,B1,C3,D1	3/2	3	[A2,C3] [A2,D1] [C3,D1]
T3	A1,B1,C2,D1	4//1	2	[A1,C2] [C2,D1]
T8	A2,B1,C1,D2	4/2	2	[A2,C1] [C1,D2]

TABLE 4 GENERATED TEST SUITE FOR 3-WAY COMBINATORIAL INTERACTION.

Test Case No.	Test Case	Iteration/ Child Thread No.	Max Weight	Covered pairs
T1	A1,B1,C1,D1	1/1	4	[A1,B1,C1][A1,B1,D1][A1,C1,D1][B1,C1,D1]
T4	A1,B1,C2,D2	1/1	4	[A1,B1,C2][A1,B1,D2][A1,C2,D2][B1,C2,D2]
T8	A2,B1,C1,D2	½	4	[A2,B1,C1][A2,B1,D2][A2,C1,D2][B1,C1,D2]
T9	A2,B1,C2,D1	½	4	[A2,B1,C2][A2,B1,D1][A2,C2,D1][B1,C2,D1]
T5	A1,B1,C3,D1	2/1	3	[A1,B1,C3][A1,C3,D1][B1,C3,D1]
T12	A2,B1,C3,D2	2/1	3	[A2,B1,C3][A2,C3,D2][B1,C3,D2]
T2	A1,B1,C1,D2	3/1	1	[A1,C1,D2]
T3	A1,B1,C2,D1	3/1	1	[A1,C2,D1]
T6	A1,B1,C3,D2	3/1	1	[A1,C3,D2]
T7	A2,B1,C1,D1	3/2	1	[A2,C1,D1]
T10	A2,B1,C2,D2	3/2	1	[A2,C2,D2]
T11	A2,B1,C3,D1	3/2	1	[A2,C3,D1]

```

Begin
    {For the remaining parameters the execution takes place in parallel}

    For parameters Pj, j=2,3,.....N do
        Where N is the total number of parameters

        Begin
            For each Test (Vi1, Vi2,.....Vim) in Ti do
                Where i = 1,2,.....X, X is the number of values of parameter p1 and m is the maximum number of test cases in list Ti at that Time

                Begin
                    Replicate the Test as many times as (the number of values of Pj – 1)
                    Add all the replicated nodes sequentially after the current original test node and before the other test nodes in Ti
                    For each value in Pj do
                        Begin

                            Append the original node with V1 and all the replicated tests with (V2, V3,.....Vy-1, Vy) where Vy is a value of Pj and each of which is considered in order.

                        End

                    End
                End
            End
        End
    End
End
    
```

B. Tree generation Algorithm for Child Thread

The tree generation algorithm thus provides the following advantages:

1. A systematic method whereby all possible test cases are generated in order.
2. The above procedure works fine with the parameters having any number of values. Therefore all parameters can have different or same values as any real time system to be tested would have.
3. The procedure appears to generate the full tree by using all the values of the parameters but at every iteration only a set of leaf nodes are left thus having a list of leaf nodes (or test cases) when the procedure ends.
4. Since the test cases in every branch are generated in parallel by the child threads there is significant reduction in time.

The example tree shown in Fig. 1 explains how the test cases are constructed manually. In reality we may need only the leaf nodes and all the intermediate nodes are not used. Therefore in order to increase the efficiency of the implementation we have constructed the same tree as in Fig. 1 using the proposed parallel tree generation algorithm. This proposed algorithm constructs the tree by minimising the number of nodes. Minimisation of the number of nodes is

achieved by giving importance only to the leaf nodes at every stage. The main thread just constructs the base branches of the tree each of which consists of one value of the first parameter in an order in which the input was made. Therefore, in the example above there are only two base branches and the value A₁ is assigned to branch T₁ and A₂ to T₂. Then the main algorithm invokes a number of unique child threads to handle each of the branches separately. At each stage or iteration each of the child threads look at the leaf nodes of their corresponding branches and generate the next level nodes by considering all the values of the current parameter, to generate the new set of nodes. The new set of leaf nodes from an already existing set is calculated using a replication strategy. The existing set of leaf nodes be E_{soln}, new set of leaf nodes be N_{soln} and the number of values of the parameter under consideration be n. Then,

$$N_{soln} = E_{soln} * n \tag{1}$$

Let there be 4 leaf nodes in a branch and the next parameter to be considered has 2 values. Then the new list of nodes for that branch will have 8 new leaf nodes as a result. The algorithm considers every leaf node separately and calculates the number of times this particular node needs to be replicated with the formulae given below:

$$\text{The number of values of } p_j - 1 \quad (2) \quad (2)$$

Where $p_j -$ is the j^{th} parameter under consideration for constructing the new set of leaf nodes and $j=1, 2, \dots, N -$ the number of parameters. In the Fig. 1 that is shown above consider the leaf nodes (A1, B1) of list or branch T1 and (A2, B1) of branch T2. To construct the next level of leaf nodes the parameter under consideration is C, which has values C1, C2 and C3. Therefore, the node (A1, B1) needs to be replicated twice. Now we will have three (A1, B1) nodes to which C1 is added to the first, C2 is added to the second and C3 is added to the third and then the replicated nodes are included in the list of leaf nodes after the original node. The same is done to (A2, B1). It is replicated twice and hence we have three of it (one original and two replicated nodes). Now C1 is added to the first (original node), C2 is added to the second (replicated node) and C3 is added to the third (replicated node). Thus we have (A2, B1, C1), (A2, B1, C2) and (A2, B1, C3). If there are more parameters the same is continued until all the parameters are considered. Thus, once the lists of leaf nodes are generated we go to the next strategy of iterative and parallel cost calculation to construct the test suite.

IV. TEST SUITE GENERATION BY ITERATIVE AND PARALLEL COST CALCULATION STRATEGY

The main thread includes the base test cases which would definitely have a maximum cost value and then invokes a number of unique child threads which operate in parallel on each of the branches lists. The main thread iterates $N-2$ times thus generating $N-2$ test suites. In the first iteration, $i=2$, the child threads iterate through the lists of test cases until all the pairs of the 2-way covering array are covered. Then the minimum 2-way test suite generated is stored and the next iteration begins. Now, $i=3$ and the child threads iterates again through the lists of test cases until all the 3-way combinations of the 3-way covering array are covered and then the 3-way test suite generated is stored. Thus this is continued until $i= N-1$. At each iteration, all the test cases with the maximum cost (W_{max}) for that particular iteration are included in the test suite. Thus the algorithm guarantees identifying minimum test suites for parameters with same as well as different number of values.

A. Strategy T-way Test Suite Generation by Iterative and Parallel Cost Calculation (Main Thread)

Input: Lists of test cases. Each list holds the test cases generated by the tree in one particular branch of that tree.

Output: T-way test suites with minimum number of test cases

Begin

Temp_b = T_b (where b is the number of lists of test cases)

X = number of values of parameter p1

B = min (Value(p1), Value(p2),Value(pn))

For i = 2 to N-1 do

Begin

Generate the i-way covering array for the given parameters.

$w_{max} = N! / ((i!) * ((N-i)!))$ // N – is the number of parameters

Let T' be an empty set where i-way test suites are stored.

For a = 1 to B do

Begin

Test_a = concatenate the ath values of all the parameters to form a test case.

End

For each Test_a do

Begin

Delete all the T-way combinations that Test_a covers in the covering array

Delete Test_a from the T_i Lists

T' = Test_a

End

Creates a set of temporary lists Y_i corresponding to the T_i lists, where i= 1,2,.....X, X is the number of values of parameter p1 or the number of lists.

Create X threads with unique thread ids. Assign every child thread Th_i with one T_i lists, the corresponding Y_i lists, i value and W_{max} value, and execute all the child threads in parallel.

Wait for the termination of all the child threads.

Store the i-way test suite generated in the list T'

T_b = Temp_b

End

End

B. Strategy T-way Test Suite Generation by Iterative and

```
Begin
While (covering array is not empty) do
Begin
  For each Test  $T_{ij}$  in  $T_i$  do
    Where  $i=1,2,\dots,X$ ,  $X$  – is the number of lists and  $j=1,2,\dots,n$  where there are  $n$  test cases in  $T_i$  at that time
    Begin
      Cost[ $T_{ij}$ ]= The number of T-way combinations covered by it in the covering array
      If (Cost[ $T_{ij}$ ]==Wmax)
        Begin
           $Y_i = T_{ij}$ 
        End
      End
    End
  End
  {Whichever thread completes its execution first locks the covering array and updates all its test cases with Wmax values from  $Y_i$  to the Test suite  $T'$  and deletes all the corresponding T-way combinations of those test cases included in  $T'$  from the covering array. The other threads on completing execution enters a queue and does its updation in that queued order by locking and unlocking the covering array after the first thread releases its lock on the covering array }
  For each  $Y_i$  do (lock the covering array and make updation)
    Begin
      If ( $Y_i \neq$  empty)
        Begin
          For each Test  $T_{ij}$  in  $Y_i$  do
            Begin
              Count= The number of T-way combinations covered by it in the covering array
              If (Count ==Wmax)
                Begin
                   $T' = T' \cup T_{ij}$ 
                  Delete all the T-way combinations that  $T_{ij}$  covers in the covering array
                  Delete  $T_{ij}$  from the lists  $T_i$ 
                End
              End
            End
          Delete  $T_{ij}$  from the lists  $Y_i$ 
        End
      End
    End
  End
  (unlock the covering array)
  End
  Wait until all child threads finishes updating
  Wmax=Wmax-1
End
End
```

Parallel Cost Calculation (Child Thread)

V. CONCLUSION

In this paper we have explained in details the parallel tree based test data generation and parallel iterative cost calculation strategy for multi-way combinatorial interaction testing and the correctness of the proposed strategy has been proved in section 3 (Tables 1, 2, 3 and 4).

REFERENCES

- [1] M. F. J. Klaib, K. Z. Zamli, N. A. M. Isa, M. I. Younis, R. Abdullah, "G2Way – A Backtracking Strategy for Pairwise Test Data Generation", in the 15th IEEE Asia-Pacific Software Engineering Conference, Beijing, China, 2008, pp. 463-470.
- [2] D. M. Cohen, S. R. Dalal, M. L. Fredman, G. C. Patton, "The AETG System: An Approach to Testing Based on Combinatorial Design", in IEEE Transactions on Software Engineering, vol. 23, 1997, pp. 437-444.
- [3] Y. Lei, R. Kacker, D. Kuhn, V. Okun, J. Lawrence, "IPOG/IPOD: Efficient Test Generation for Multi-Way Software Testing", in Journal of Software Testing, Verification, and Reliability, vol. 18, 2009, pp.125-148.
- [4] M. B. Cohen, "Designing Test Suites for Software Interaction Testing", in Computer Science, University of Auckland, Ph.D, New Zealand, 2004.
- [5] D. M. Cohen, S. R. Dalal, A. Kajla, G. C. Patton, "The Automatic Efficient Test Generator (AETG) System", in the 5th International Symposium on Software Reliability Engineering, Monterey, CA, USA, 1994, pp. 303-309.
- [6] Y. Lei, K. C. Tai, "In-Parameter-Order: A Test Generation Strategy for Pairwise Testing", in the 3rd IEEE International. High-Assurance Systems Engineering Symp, Washington, DC, USA, 1998, pp. 254-261.
- [7] T. Shiba, T. Tsuchiya, T. Kikuno, "Using Artificial Life Techniques to Generate Test Cases for Combinatorial Testing", in the 28th Annual International Computer Software and Applications Conf. (COMPSAC'04), Hong Kong, 2004, pp. 72-77.
- [8] K. C. Tai, Y. Lei, "A Test Generation Strategy for Pairwise Testing", in IEEE Transactions on Software Engineering, vol. 28, 2002, pp. 109-111.
- [9] S. R. Dalal, A. Jain, N. Karunanithi, J. M. Leaton, C. M. Lott, G. C. Patton, B. M. Horowitz, "Model Based Testing in Practice", in the International Conf. on Software Engineering (ICSE), 1999, pp. 285-294.
- [10] D. R. Kuhn, M. J. Reilly, "An Investigation of the Applicability of Design of Experiments to Software Testing", in the 27th NASA/IEEE Software Engineering Workshop, 2002, pp. 69-80.
- [11] D. R. Kuhn, V. Okun, "Pseudo-Exhaustive Testing for Software", in: the 30th Annual IEEE/NASA Software Engineering Workshop (SEW '06), 2006, pp. 25-27.
- [12] D. R. Kuhn, D. R. Wallace, A. M. Gallo, "Software Fault Interactions and Implications for Software Testing", in IEEE Transactions on Software Engineering vol. 30, June 2004, pp. 418-421.
- [13] J. Yan, J. Zhang, "A Backtracking Search Tool for Constructing Combinatorial Test Suites", in Journal of Systems and Software - Elsevier, vol. 81, October 2008 pp. 1681-1693.
- [14] R. Bryce, C. J. Colbourn, "Prioritized Interaction Testing for Pairwise Coverage with Seeding and Avoids", in Information and Software Technology Journal (IST, Elsevier), vol. 48, October 2006, p. 960-970.
- [15] D. R. Kuhn, Y. Lei, R. Kacker, "Practical Combinatorial Testing: Beyond Pairwise", in IT Professional- IEEE Computer Society vol. 10, May 2008, pp. 19-23.
- [16] D. A. Bader, W. E. Hart, C. A. Phillips, "Parallel Algorithm Design for branch and bound", in Tutorials on Emerging Methodologies and Applications in Operations Research. Mathematics and Statistics, vol. 76. Springer, New York , 2005 pp. 5.1- 5.44.
- [17] R. Setia, A. Nedunchezhiyan, S. Balachandran, "A New Parallel Algorithm for Minimum Spanning Tree Problem". in the 16th Annual IEEE International Conference on High Performance Computing. Cochin, India, 2009, pp 1-25.
- [18] Kamal Z. Zamli, Mohammad F.J. Klaib, Mohammed I. Younis, Nor Ashidi Mat Isa, and Rusli Abdullah, "Design and implementation of a t-way test data generation strategy with automated execution tool support" Information Sciences, vol 181, issue 9, May 2011, pp 1741-1758.
- [19] Mohammad F. J. Klaib, Sangeetha Muthuraman, A. Noraziah, "A Tree Based Strategy for Interaction Testing", in The 5th International Conference on Information Technology 2011 (ICIT2011), AL-Zaytoonah University of Jordan, Faculty of Science & Information Technology, Jordan, , May, 2011, pp 1-5.
- [20] Y. Lei, R. Kacker, D. R. Kuhn, V. Okun, J. Lawrence, IPOG/IPOG-D: efficient test generation for multi-way combinatorial testing, Software Testing, Verification and Reliability, vol 18, Issue 3, September 2008 pp 125-148.
- [21] Z. Hisham C. Soh, M. I. Younis, S. Abdullah, K. Zamli, "Distributed t-way Test Suite Generation Algorithm for Combinatorial Interaction Testing", in the International conference on IT to Celebrate S. Charmonnan's 72nd Birthday (Charm09), Thailand, March 2009, pp. 431-437
- [22] M. I. Younis, K. Z. Zamli, "MC-MIPOG: A Parallel t-Way Test Generation Strategy for Multicore Systems", in ETRI Journal, vol. 32, no. 1, February 2010 pp. 73-83..
- [23] M. Grindal, B. Lindstrom, J. Offutt, S. F. Andler, "An Evaluation of Combination Strategies for Test Case Selection". Technical Report HS-IDA-TR-03-001, Department of Computer Science, University of Skövde, 2003.
- [24] M. F. Klaib, S. Muthuraman, N. Ahmad, and R. Sidek, "Tree Based Test Case Generation and Cost Calculation Strategy for Uniform Parametric Pairwise Testing", in Journal of Computer Science, Journal of Computer Science, 6 (4), 2010, pp: 425-430.
- [25] M. F. J. Klaib, S. Muthuraman, N. Ahmad, and R. Sidek, "A Tree Based Strategy for Test Data Generation and Cost Calculation for Uniform and Non-Uniform Parametric Values", in International Symposium on Frontier of Computer Science, Engineering and Applications (CSEA2010), Bradford, UK, 2010, pp 1376 - 1383.
- [26] M. F. J. Klaib, S. Muthuraman, N. Ahmad, and R. Sidek, "A Parallel Tree Based Strategy for Test Data Generation and Cost Calculation for Pairwise Combinatorial Interaction Testing", in The Second International Conference on Networked Digital Technologies (NDT2010) Charles University in Prague, Czech Republic: Springer, 2010, pp 509-522.
- [27] D.R. Kuhn, R.N. Kacker and Y. Lei, "Combinatorial Coverage as an Aspect of Test Quality", Journal of Defense Software Engineering, 2014.
- [28] D.R. Kuhn, R.N. Kacker and Y. Lei, "Measuring and Specifying Combinatorial Coverage of Test Input Configurations", Innovations in Systems and Software Engineering: a NASA journal, 2014, pp1-15.
- [29] J. Torres-Jimenez, I. Izquierdo-Marquez, "Survey of Covering Arrays", in the 15th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC 2013), Timisoara, Romania, 23-26, 2013, pp. 20-27.
- [30] R.N. Kacker, D.R. Kuhn, Y. Lei, and J.F. Lawrence, "Combinatorial Testing for Software, an Adaptation of Design of Experiments", Measurement, vol. 46, no. 9, 2013, pp. 3745-3752.
- [31] X. Niu, C. Nie, Y. Lei, A.T.S. Chan, "Identifying Failure-Inducing Combinations Using Tuple Relationships", in the 6th IEEE International Conference on Software, Testing, Verification and Validation (ICST 2013), Luxembourg, March 18-22, 2013, pp. 271-280.
- [32] M.N. Borazjany, L.S.G. Ghandehari, Y. Lei, R.N. Kacker and D.R. Kuhn, "An Input Space Modeling Methodology for Combinatorial Testing", in the 6th IEEE International Conference on Software, Testing, Verification and Validation (ICST 2013), Luxembourg, March 18-22, 2013, pp. 372-381.

Software Engineering

Using MADA+TOKAN to Generate Use Case Models from Arabic User Requirements in a Semi-Automated Approach

Nabil Arman

Department of Computer Science and Engineering
Palestine Polytechnic University
Hebron, Palestine
narman@ppu.edu

Abstract—Automated software engineering has attracted a large amount of research efforts. The need for new approaches that reduces the cost of developing software systems within project schedule has made it necessary to develop approaches that aid in the construction of different UML models in a semi-automated approach from Arabic textual user requirements. UML use case models represent an essential artifact that provide a perspective of the system under analysis or development. The development of such use case models is very crucial in an object-oriented development methodology. In this paper, MADA TOKAN is used to parse different statements of the user requirements written in Arabic to obtain different components of a sentence like lists of nouns, noun phrases, verbs, verb phrases, etc. that aid in finding potential actors and use cases. A set of steps that represent our approach for constructing a use case model is presented. Finally, the proposed approach is to be validated and implemented at a later stage of the research project.

Keywords— *Arabic User Requirements, Use Case Model, MADA+TOKAN tool.*

I. INTRODUCTION

Object-oriented methodologies are used for software systems development for the many benefits they provide like software reuse, reducing software development costs, to name just a few. Therefore, there is a need for development of automated tools that can help in constructing different components of an object-oriented software system.

A use case diagram shows a set of use cases and actors and their relationships. Use case diagrams address the static use case view of a system. These diagrams are especially important in organizing and modeling the behaviors of a system. This paper addresses the problem of generating a use case model from user requirements, written in Arabic, in a semi-automated approach. An Arabic natural language processing tool/software, namely MADA+TOKAN, is used to parse different statements of the user requirements, written in Arabic, to obtain lists of nouns, noun phrases, verbs, verb phrases, etc. that aid in finding potential actors and use cases. A set of steps that represent our approach for constructing a use case model is presented.

The rest of the paper is organized as follows: the section about related works presents the literature review and any related approaches; the section about constructing use cases describes the process of constructing use case models from Arabic user requirements; the section about validation presents the validation and implementation of our proposed approach, and finally, the section about conclusion presents the main issues related to the proposed approach.

II. RELATED WORKS

Recently there is a great interest in automating software engineering activities. Many tools were developed to automate different activities of software systems development like normalizing relational database schemas, reverse engineering of relational database and generating the corresponding entity-relationship data model, ...etc. [1, 2]. In addition, many CASE tools were developed to aid in drawing different diagrams of UML. For example, Rational Rose is an object-oriented Unified Modeling Language (UML) software design tool intended for visual modeling and component construction of enterprise-level software applications [3]. Rational Unified Process (RUP) is an object-oriented Web-enabled program development methodology. [4].

More advanced tools were developed to automate software engineering activities that are more complicated than just aiding in drawing a UML diagram or checking its overall structure. Arman and Daghameen proposed a systematic approach that generates class diagrams from textual software requirements. They presented some steps to build a matrix that was used to obtain classes and their associations to generate class diagrams [5]. The same authors later developed a CASE tool, called, SDLCCASE tool that implemented their approach [6]. Kothari proposed an approach that can extract the basic elements for generating a class diagram from user requirements written in a clear way. The Natural Language Processing for Class (NLPC) can extract classes, data members and member functions from the given user requirements [7]. This approach was implemented

as a software tool to generate the class diagrams. Seresht and Ormandjieva proposed an approach to generate use case diagrams from software requirements, but this approach depends on other models to obtain the use case by combining two technologies: Recursive Object Model (ROM) and Expert Comparable Contextual (ECC) Models and it doesn't deal with the textual requirements directly [8]. Cayaba et al. proposed an approach called computer automated use case diagram generator (CAUse), that can generate the use case diagrams from a text described using a special language called ADD [9]. However, this approach depends on the ADD language to generate the use case diagrams. Mala and Uma proposed an approach to extract the object-oriented elements of system requirements. This approach started by assigning the parts of speech tags to each word in the given requirements [10]. An automated approach that helps a software engineer in developing formal specifications in VDM is presented in [11]. In this approach, the detection of ambiguous sentences and inconsistencies in the informal specifications was a major concern. Relationships are determined from the verbs in the sentences. Entities and relationships are then used to develop an entity-relationship model from which a VDM data types are obtained. Another major research endeavor in automated software engineering was the work of using natural language processing to aid in object-oriented analysis [12]. The natural language processing capabilities to build a UML class diagram was used. The research approach involved two major stages. The first stage is a linguistic analysis of the text to build a semantic net. The second stage uses the semantic net to obtain the class model and its (classes, associations, attributes, etc.). Arman and Jabbarin used Stanford Parser to construct uses cases from Arabic user requirements [13]. In this paper, MADA+TOKAN is used since it provides a richer set of tags that can help in parsing the Arabic statements more accurately. In addition, more accurate heuristics are presented.

III. CONSTRUCTING USE CASE MODELS

The IEEE, in the standard for Software Requirements Specification, identifies a good requirement as correct, unambiguous, verifiable, and traceable. The IEEE also identifies a good set of requirements as complete, consistent, and modifiable. This is an assumption that is used in our approach. It is assumed that the requirements are "good" in the sense implied by the IEEE good requirements assumptions.

This section describes how the actors and use cases are extracted from user requirements written in Arabic. There is a need for an Arabic Natural Language Processing tool such as the MADA+TOKAN, which is used in this research to help in splitting and tokenizing the Arabic user requirements text. Once this is performed, a set of heuristics are used to construct the use case model as presented in subsequent subsections.

A. MADA+TOKAN

MADA+TOKAN is a versatile, highly customizable and freely available toolkit for Arabic NLP applications. It consists of two components. MADA is a utility that, given raw Arabic text, adds as much lexical and morphological information as

possible by disambiguating in one operation part-of-speech tags, lexemes, diacritizations and full morphological analyses. TOKAN is a utility that, given the information MADA produces, can generate a tokenization (sometimes also called a "segmentation") formatted exactly to user specifications. This tokenization also identifies the stem of the word [14]. All user requirements are processed using the MADA+TOKAN.

A set of user requirements for a system implementing ridesharing is used. The requirements were written in Arabic and some of these requirements are used in our examples. The ridesharing system includes many requirements. Two examples are presented below:

- يقوم السائق بتسجيل الدخول الى النظام ومن ثم يستطيع الاعلان عن الرحلة التي سيقوم بها ويقوم في هذه المرحلة بتحديد ومتطلباتها وتشمل: وقت الرحلة (الانطلاق) و المسار الذي سيسلكه اضافة الى عدد المقاعد الفارغه. كما ويستطيع حذف رحلة بعد انتهاءها او الغائها.

A translation of the examples: "The driver shall be able to sign in to the system and then he shall be able to make an advertisement about the trip he is going to make. At this stage, he provides all information related to the trip, including the time and the number of seats available. He shall also be able to delete the trip afterwards."

- يقوم السائق بقبول الركاب او رفضهم , يستطيع ايضا تتبع المسافرين باستخدام ال GPS ان توفرت هذه الخاصية عند الركاب في النهاية يقوم بتسجيل الخروج.

A translation of this example: "The driver shall be able to accept or reject passengers. He shall also be able to follow the passengers using a GPS if available. At the end, he shall also be able to sign out."

In addition, MADA+TOKAN uses a set of tags to describe different components of a statement.

MADA+TOKAN tokenizes the statements and uses a large number of tags, including:

MADA+TOKAN has many tags, including, but not limited to:

Verb : VBP, VBZ, VBD

Noun: DTNN , NN ,DTNNS ,NNS ,DTNNP ,NNP ,DTNNPS ,NNPS

Addictive (Object): DTJJ, JJ

Preposition: IN

Connectors: CC (و/ف،أو) , RB ثم

These tags are used in determining the actors and uses cases as described below.

B. Actors Identification

To identify the actors from the user requirements written in Arabic, a set of heuristics are presented. These heuristics are used to extract the actors from the tagging of the user requirements generated from the MADA+TOKAN. These heuristics are presented as follows:

- If the statement is simple (i.e. it contains only a verb, a subject and an object) then the actor is the main subject in the statement.

e.g. يقوم السائق بتسجيل الدخول

Using MADA+TOKAN tags, the statement is divided into:

يقوم/VBP السائق/DTNN ب/IN تسجيل/NN الدخول/DTNN

Here the main subject is السائق and it's the actor.

Generalization: If the statement is in the form of <VBP> <DTNN> <IN> <NN> <DTNN> when using the MADA+TOKAN, then the first DTNN is the actor. To simplify referencing, the form can be write as <VBP> <DTNN₍₁₎₍₂₎

- When there are two statements combined with a connection then, there are three cases:

a) The subject is the actor.

e.g. يقوم السائق بتسجيل الدخول إلى النظام و من ثم يستطيع الإعلان عن الرحلة

Using MADA+TOKAN tags, the statement is divided into:

يقوم/VBP السائق/DTNN ب/IN تسجيل/NN الدخول/DTNN
يستطيع/VBP الإعلان/DTNN عن/IN الرحلة/DTNN
و/CC من/WP ثم/NN إلى النظام/DTNN

The actor is السائق.

b) If the subject is redundant in the second statement then the actor doesn't change.

e.g. يقوم السائق بتسجيل الدخول إلى النظام و من ثم يستطيع الإعلان عن الرحلة

Using MADA+TOKAN tags, the statement is divided into:

يقوم/VBP السائق/DTNN ب/IN تسجيل/NN الدخول/DTNN
يستطيع/VBP الإعلان/DTNN عن/IN الرحلة/DTNN
و/CC من/WP ثم/NN إلى النظام/DTNN

The actor is السائق.

c) If the subject changes in the second statement then this is another actor.

e.g. يقوم السائق بتسجيل الدخول إلى النظام و من ثم يستطيع الراكب اختيار الرحلة المعطن عنها من خلال زيارة النظام.

Using MADA+TOKAN tags, the statement is divided into:

يقوم/VBP السائق/DTNN ب/IN تسجيل/NN الدخول/DTNN
يستطيع/VBP الإعلان/DTNN عن/IN الرحلة/DTNN
و/CC من/WP ثم/NN إلى النظام/DTNN
و/CC من/WP ثم/NN إلى النظام/DTNN
و/CC من/WP ثم/NN إلى النظام/DTNN

The actors are السائق and الراكب .

Generalization: If the statement is in the form of <VBP₍₁₎₍₁₎₍₁₎₍₂₎₍₁₎₍₃₎₍₁₎₍₁₎₍₂₎₍₄₎₍₂₎₍₅₎₍₁₎₍₁₎₍₁₎₍₂₎₍₂₎₍₃₎₍₁₎₍₆₎₍₃₎₍₄₎₍₅₎₍₂₎₍₁₎₍₇₎₍₁₎₍₃₎₍₈₎₍₂₎₍₃₎₍₃₎₍₄₎₍₄₎₍₉₎₍₂₎₍₄₎₍₂₎₍₅₎₍₆₎₍₇₎₍₈₎₍₅₎₍₉₎

C. Use Cases Identification

To identify the use cases from the user requirements, more heuristics that can be used to extract the use cases from the user requirements are presented.

- If the statement is simple (i.e. it contains only a verb, a subject and an object) then the use case is the main object in the statement.

e.g. يقوم السائق بتسجيل الدخول

Using MADA+TOKAN tags, the statement is divided into:

يقوم/VBP السائق/DTNN ب/IN تسجيل/NN الدخول/DTNN

The main object is تسجيل الدخول and it's the use case.

Generalization: If the statement is in the form of <VBP> <DTNN₍₁₎₍₂₎₍₂₎ is the use case.

- If the statement contains the connector (و) without any verb or actor in the second statement then the second statement is the use case.

e.g. يستطيع الراكب الانضمام إلى الرحلة و الحجز فيها

Using MADA+TOKAN tags, the statement is divided into:

يستطيع/VBP الراكب/DTNN الانضمام/DTNN إلى/IN الرحلة/DTNN
و/CC الحجز/DTNN في/IN

In the above example, the statement contains a connector (و) so there are two use cases 1- يستطيع الانضمام 2- يستطيع الحجز

Generalization: If the statement is in the form of <VBP> <DTNN₍₁₎₍₂₎₍₁₎₍₃₎₍₄₎₍₂₎₍₅₎

- a) The use case is the VBP with DTNN₍₂₎.
b) The use case is the VBP with the DTNN₍₄₎ after the CC.
• If the statements that contain the connector (أو) without any verb or actor in the statement then the first verb in the statement with the first noun after each connector is a use case.

e.g. يستطيع الراكب الانضمام إلى الرحلة و الحجز فيها أو الانسحاب منها.

Using MADA+TOKAN tags, the statement is divided into:

يستطيع/VBP الراكب /DTNN الانضمام /DTNN إلى /IN الرحلة /DTNN و /CC الحجز /DTNN في /IN ها /PRP ما /PRP من /DTNN الانسحاب

In the above example, the statement contains a connector (e.g. أو) so there are three use cases 1- يستطيع الانضمام 2- يستطيع الحجز 3- يستطيع الانسحاب.

Generalization: If the statement is in the form of <VBP> <DTNN₍₁₎> <DTNN₍₂₎> <IN₍₁₎> <DTNN₍₃₎> <CC₍₁₎> <DTNN₍₄₎> <IN₍₂₎> <PRP₍₁₎> <CC₍₂₎> <DTNN₍₅₎> <WP> <PRP₍₂₎> when using the MADA+TOKAN, then

- a) The use case is the VBP with DTNN₍₂₎.
b) The use case is the VBP with DTNN₍₄₎ after the CC₍₁₎.
c) The use case is the VBP with DTNN₍₅₎ after the CC₍₂₎.

D. Use Case Model Generation

To complete the generation of the use case model, a structure that depicts the relationships among the different tokens is needed. A matrix consisting of columns with headings, which contain the potential use cases, and rows with labels, which contain the potential actors, is used. These are obtained from the heuristics explained previously. The matrix is filled by arrow symbols. An arrow means that an actor is associated with one or more particular use cases. For example, if an arrow is shown in the cell that corresponds to the row labeled with Use Case i and the column labeled with Actor j, it is concluded that Actor j is associated with Use Case i.

Once the matrix is constructed, the use case model is obtained by taking an actor with all its associated use cases to generate a use case diagram. The set of all use case diagrams represent the use case model. According to the above description, this approach can be implemented easily to generate a use case model.

Applying the proposed approach described so far to the set of user requirements mentioned previously generates the matrix presented in Table I.

As can be seen from the table, all potential actors are associated with the related use cases using the arrow notation.

TABLE I. MATRIX OF POTENTIAL ACTORS AND THEIR USE CASES

Potential Actors \ Potential Use Case	الراكب	السائق	المدير
تسجيل الدخول	←	←	←
يستطيع الإعلان		←	←
تحديد متطلبات		←	
قبول الراكب		←	
ننبح المسافرين		←	
تسجيل الخروج	←	←	←
يستطيع الانضمام	←		
يستطيع الحجز	←		
يستطيع الانسحاب	←		
يوفر تغذية	←		
يستطيع اضافة			←
يستطيع حذف			←
تصنيف مستخدمين			←
يستعرض الرحلات	←		←

IV. PROPOSED APPROACH VALIDATION AND IMPLEMENTATION

The next step in this research is to validate the proposed approach. Once the approach proves to be beneficial, the proposed approach will be implemented as a software tool that can be used to generate the use case model from Arabic user requirements.

V. CONCLUSIONS

The proposed approach of developing use case models is very essential in the practice of object-oriented software engineering. This approach can be implemented and incorporated in any Integrated CASE (Computer Aided Software engineering) Tool to aid in the process of obtaining the use case models from user requirements written in Arabic. The approach has the main advantage of dealing with Arabic language. In addition, a set of heuristics are presented to obtain the use cases. These heuristics use the tokens produced by a natural language processing tool, namely MADA+TOKAN. These tokens are then used as the main components of the use case diagram, namely, the actors and the use cases. Finally, the proposed approach is to be validated and implemented in further research efforts.

ACKNOWLEDGMENT

The author would like to thank the Software Engineering Research Group members at Palestine Polytechnic University, especially Mr. Ibrahim Nassar for his help regarding

MADA+TOKAN tool and Dr. Dia Abu Zeineh for his help regarding the use of Natural Languages Processing tools.

REFERENCES

- [1] N. Arman, "Normalizer: A Case Tool to Normalize Relational Database Schemas, *Information Technology Journal*, pp. 329-331, Vol. 5, No. 2, ISSN: 1812-5638, 2006.
- [2] N. Arman, "Towards E-CASE Tools for Software Engineering," *International Journal of Advanced Corporate Learning*, pp. 16-19, Vol. 6, No. 1, 2013.
- [3] <http://searchciomidmarket.techtarget.com/home/0,289692,sid183,00.html>, accessed: October 15, 2013.
- [4] "Rational Unified Process (RUP)": ch1, Prentice Hall 1990, ISBN 0-13-629841-9.
- [5] N. Arman. and K. Daghameen, "A Systematic Approach for Constructing Static Class Diagrams from Software Requirements," *International Arab Conference on Information Technology (ACIT2007)*, November 26-28 2007, Amman, Jordan.
- [6] K. Daghameen and N. Arman. "Requirements Based Static Class Diagram Constructor (SCDC) CASE TOOL." *Journal of Theoretical & Applied Information Technology*, pp. 108-114, Vol15, No. 2, 2010.
- [7] P. Kothari, "Processing Natural Language Requirement to Extract Basic Elements of a Class," *International Journal of Applied Information Systems (IJ AIS)*, ISSN : 2249-0868.
- [8] S. Seresht and O. Ormandjieva, "Automated Assistance for Use Cases Elicitation from User Requirements Text," 11th. Workshop on Requirement Engineering, 2009.
- [9] C. Cayaba, J. Rodil and N. Lim, "CAUse: Computer Automated Use Case Diagram Generator", 2006.
- [10] G. Mala and G. Uma, "Automatic Construction of Object Oriented Design Models [UML Diagrams] from Natural Language Requirements Specification", 2006.
- [11] F. Meziane, "From English to Formal Specification", PhD thesis Dept. of Maths and Computer Science, University of Salford, UK, 1994.
- [12] H. Harmain, "Building Object-Oriented Conceptual Models Using Natural Language Techniques", PhD thesis, University of Sheffield, 2000.
- [13] N. Arman and S. Jabbarin, "Generating Use Case Models from Arabic User Requirements in a Semiautomated Approach Using a Natural Language Processing Tool," *Journal of Intelligent Systems*, Vol. 24, No. 2, pp. 277-286, 2015.
- [14] N. Habash, O. Rambow, and R. Roth, "MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools*", pp 102-109, Cairo, Egypt, 2009.

Smart OptiSelect Preference Based Innovative Framework for User-in-the-Loop Feature Selection in Software Product Lines

Ahmed Eid El Yamany

College of Computing and Information Technology
Arab Academy for Science, Technology, and Maritime Transport
Cairo, Egypt
Ahmedeid100@gmail.com

Mohamed Shaheen Elgamel

College of Computing and Information Technology
Arab Academy for Science, Technology, and Maritime Transport
Alexandria, Egypt
cshaheen@hotmail.com

Abstract—Smart OptiSelect is a multi-objective evolutionary optimization and a machine learning based framework for software product lines feature selection. It serves in the direction of filling the gap between software product lines search based feature selection optimization and real life utilization by stakeholders. OptiSelect enables system analysts and project managers to select best features to implement to meet their dynamic and always changing objectives by offering plenty of multi-objective optimized solutions that complies with these objectives. Smart OptiSelect created the availability for providing various versions of result sets based on user experience in a more comprehensive working flow. Smart OptiSelect is enabled to interactively figure out user's preferences and help to reach more convenient solutions that should best draw out the user's desires and express his organization goals.

Keywords— *User-in-the-loop (UIL); Software Product Lines; Feature Models; Optimal Feature Selection; Multi-objective Optimization; Search-Based Software Engineering; Machine Learning; Pareto Front; Non-Dominant Solutions*

I. INTRODUCTION

Smart OptiSelect is a continuous result of research experiments that investigated the best ways to empower the user in the process of feature model configuration. Two targets are achieved through this version: 1) Narrowing the gap between product lines search based optimization and real life cases to provide real utilizations to software stakeholders. 2) Provide a preference based framework which can understand the user's needs and provide effective suggestions based on them.

Smart OptiSelect is an interactive framework. Users are enabled to dynamically load feature models, apply adjustments to feature attributes, set objectives and desirable thresholds, and interact by selecting preferred solution among optimization cycles.

Smart OptiSelect is a continuing effort of the previously proposed Opti-Select [1] through enhancing the workflow using machine-learning techniques to intelligently extend preferences, hybrid multi-objective optimization, and adding new features as setting user's objective thresholds.

The optimization process takes place in an incremental form. After each round of optimization, the user is provided with a concise presentation of the multiple solutions thus make up the Pareto Front, allowing the user to mark their preferred ones to focus on producing related solutions in the following iterations.

This work discusses the features and the workflow steps of Smart OptiSelect. An overview of the used algorithms and techniques and how they work together to achieve the user's goals is provided. The rest of the paper is organized as follows; Section II illustrates Smart OptiSelect workflow steps, points of interactions with the user, and processing stages. Section III describes the algorithms and methodologies, why they are selected, and how they orchestrated to work within Smart OptiSelect. Section IV displays a survey comparing users' satisfaction with the results of different techniques. Section V summarizes the proposed framework's contributions to achieve a preference based User-in-the-Loop solutions for search based product lines features optimization. It also covers an overview of some future directions and plans.

II. SMART OPTISELECT WORKFLOW

Smart OptiSelect point of strength lays in the ability to bring together most empowered multiobjective optimization algorithms proven to produce best search based product lines features optimization results [2]. This is done side by side with

machine learning techniques in one single interface frame work giving the user the widest capability to be a part of the optimization process itself as shown in Fig. 1. This framework takes place through a tuned process to fit users' interactions.

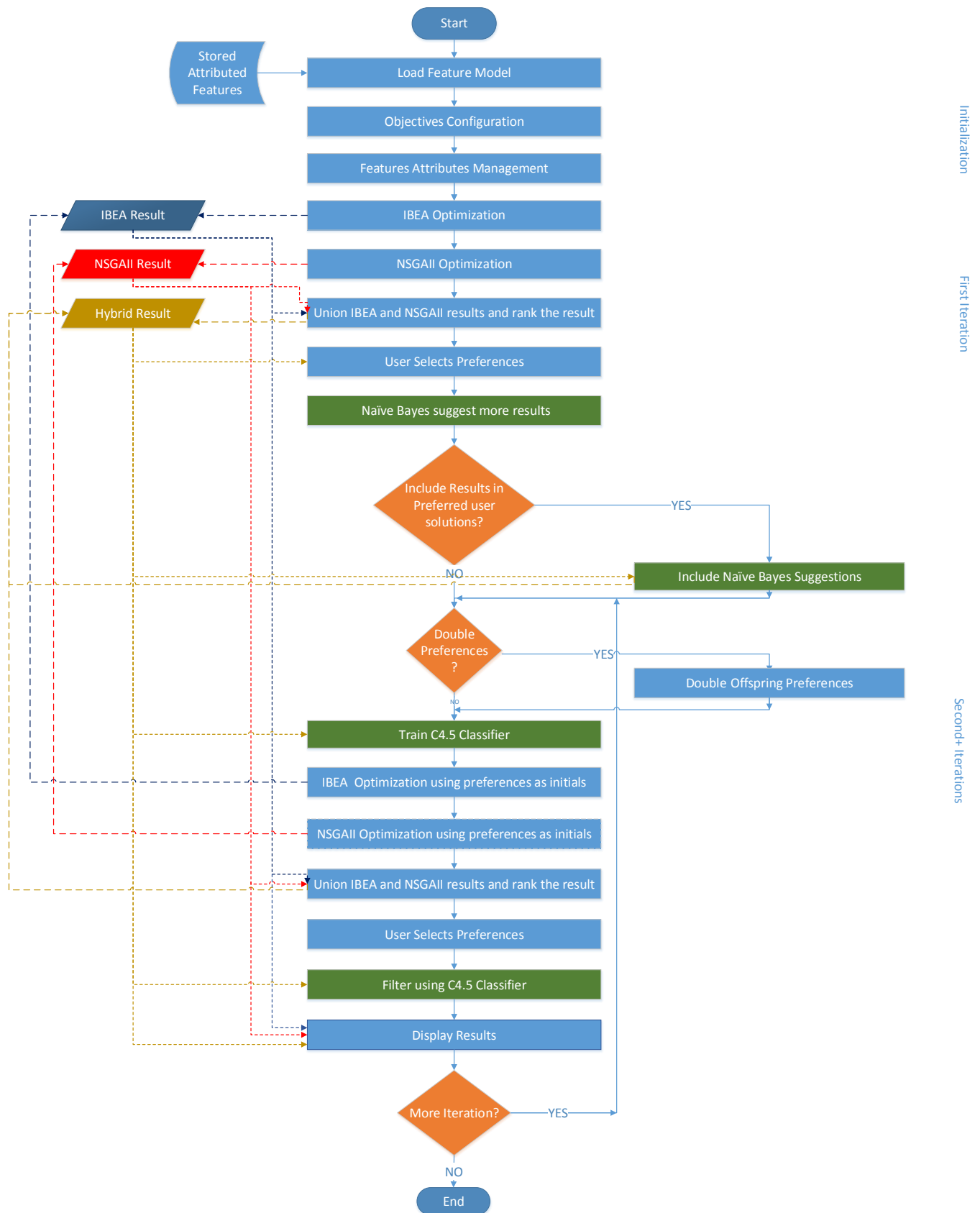


Fig. 1. Smart OptiSelect workflow diagram

A. Loading and Saving Attributed Features.

The Simple XML Feature Model (SXF) format was defined by the SPLOT website [3]. Smart OptiSelect implemented a module for dynamically reading and saving feature models in SXFM formats to decrease the time of changing the test model through configuration file or through hard coded instructions.

In Order to provide the user a capability for managing and saving changes over features' attributes. The proposed framework introduced an attributed feature model file format as shown in Fig. 2. It can attach a dynamic series of attributes to each feature in the model.

B. Objective Configuration

Smart OptiSelect has a predefined set of quality attributes for enabling the user to dynamically set optimization objectives and targets. Objectives targets are enabled through setting threshold for each objective as shown in Fig. 3.

The user is allowed to specify objectives being optimized prior to any optimization runs or between runs. This gives the user the power to use a desired solution set resulting from some objectives optimization at specific time as an offspring for a specific objective optimization.

```
ID,Desired,Excluded,UsedBefore,Cost,Deffects,Usability
web_portal,true,false,false,10.0,10,10
basic,true,false,true,16.0,3,50
html,true,false,false,20.0,3,0
```

Fig. 2. Saved faeature attributes format sample

During the optimization process, each solutions is dynamically evaluated based on the current objectives' settings by calculating their related attributes values.

C. Feature Attributes Management

Based on the selected objectives, the users are allowed to edit the corresponding attributes for each feature and define if a certain feature is forced to appear in all solutions or even to be excluded from all solutions as shown in Fig. 4.

Feature attributes management window is designed to be smart enough to help the user manage consequences of forcing existence and discarding existence of features by generating and applying corrective actions based on the behaviors of the user. It checks for user's opinion if more than one corrective option is available as shown in Fig. 5.

D. Multiobjective Optimization

Based on previous researches [4], IBEA [5] has been proven to perform better than the rest of the multiobjective algorithms in optimizing multiobjective problems related to product lines models and feature selection optimization as it pays most attention to user indicators without violating domain constraints. NSGA-II [6] came next in overall result quality.

Smart OptiSelect made advantage of both algorithms and provided innovative hybrid technique based on running both IBEA and NSGA-II separately within limited time. Then the results of both algorithms are merged employing Pareto front ranking [7].

Quality Category	Quality Attribute	Target Value	
Solution Qualities	Correctness	6	
	Features Total Count	43	
	Cost	5000	
	Defects	6	
Design Qualities	Used before	43	
	Reusability	.3	
	Conceptual Integrity	1.0	
Run-time Qualities	Maintainability	1.0	
	Performance	.2	
	Reliability	1.0	
	Scalability	1.0	
	Availability	1.0	
	Security	1.0	
	Interoperability	1.0	
	Manageability	1.0	
	User Qualities	Usability	1.0

Fig. 3. Configuration sample of the objectives being optimized

Feature	Desired	Excluded	Used Before	Defects	Cost
Web Portal(web_portal)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	10	10.0
Additional Services(a...)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	0.0
Site Statistics(site...)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	0.0
Basic(basic)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	3	16.0
Advanced(adv...)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	25.0
Site Search(site_...)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	0.0
Ad Server(ad_serv...)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	0.0
Reports(repor...)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	0.0
Pop-ups(popu...)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	2	0.0
Banners(ban...)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	0	0.0
Keyword Supp...	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	20.0
Web Server(web_se...)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	10.0
Logging(logging)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	0.0
g_id_0(id_0)[...]	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	0.0
Protocols(protocol)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	0.0
g_id_1(id_1)[...]	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	0.0
Content(cont)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	0.0
Static(static)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	50.0
Active(active)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	0.0
Persistence(persiste...)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	0.0
g_id_3(id_3)[1,1]	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	0.0
Security(r)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	10.0
g_id_4(id_4)[1,*]	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	20.0
Data Storage(d...)	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	0	300.0
Data Transfer(d...)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	1	5.0

Fig. 4. Feature attributes management window

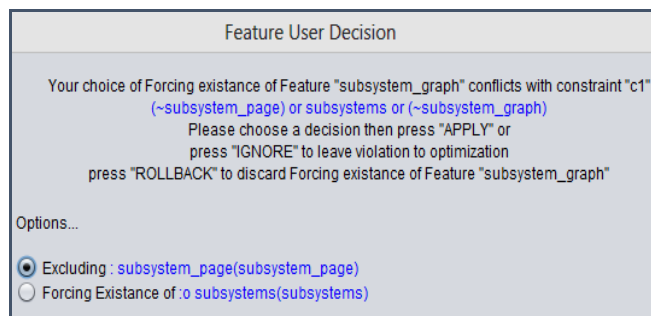


Fig. 5. Attributes management corrective actions list sample

E. User Preference Selection

Smart OptiSelect users are enabled to select a subset of the solutions from the total result set as preferred solutions as shown in Fig. 6. Selected preference are used as an initial offspring population for the next optimization cycles to force the optimization cycle to focus around the selected solutions along with the repeating cycles.

The proposed framework tries to enrich the population for the next iteration based on user selections at the current iteration. It uses any of the machine learning techniques to classify the rest of the non-selected and undisplayed solutions and see if they match the user current selections. Naïve Bayes [8] has been employed as one of the classification techniques. The user is then asked if he wants to add the suggested solutions to be considered in next iterations as shown in Fig. 7.

F. Iterating and Machine Learning

Smart OptiSelect uses final user preferred decisions selected from total result set to build and train a c4.5 classifier that aims to figure out user's preferences [9] to be used to filter result sets through next iterations.

Application repeats optimization cycle and apply user thresholds preferences filters and display different results to user to indicate if there are similar solutions to selected ones should be also selected by user.

G. Displaying Result

Smart OptiSelect provides four types of results to be displayed to the user after each iteration for comparative purposes: IBEA Result – NSGII Result – Hybrid Result – C4.5 Filtered results.

Decisions	Objectives	Will be used	Objective	Value
Decision: 0	4.0 15.0 26.0 12.0 595.0	<input checked="" type="checkbox"/>	Correctness	2.0
Decision: 1	4.0 22.0 19.0 12.0 525.0	<input checked="" type="checkbox"/>	FeatureNumber	9.0
Decision: 2	6.0 30.0 12.0 12.0 405.0	<input checked="" type="checkbox"/>	Used Before	30.0
Decision: 3	4.0 20.0 21.0 12.0 465.0	<input checked="" type="checkbox"/>	Defects	18.0
Decision: 4	6.0 29.0 13.0 12.0 415.0	<input checked="" type="checkbox"/>	Cost	686.0
Decision: 5	3.0 12.0 28.0 15.0 645.0	<input checked="" type="checkbox"/>	Reusability	0.0083333333333333
Decision: 6	4.0 23.0 18.0 12.0 445.0	<input checked="" type="checkbox"/>	Performance	0.00526315789473
Decision: 7	6.0 27.0 15.0 14.0 418.0	<input checked="" type="checkbox"/>		
Decision: 8	4.0 22.0 19.0 14.0 448.0	<input checked="" type="checkbox"/>		
Decision: 9	2.0 9.0 30.0 18.0 688.0	<input checked="" type="checkbox"/>		
Decision: 10	4.0 21.0 20.0 14.0 448.0	<input checked="" type="checkbox"/>		
Decision: 11	3.0 14.0 26.0 15.0 595.0	<input checked="" type="checkbox"/>		
Decision: 12	2.0 14.0 25.0 18.0 631.0	<input checked="" type="checkbox"/>		
Decision: 13	3.0 18.0 22.0 15.0 485.0	<input checked="" type="checkbox"/>		
Decision: 14	3.0 13.0 27.0 15.0 615.0	<input checked="" type="checkbox"/>		
Decision: 15	4.0 19.0 22.0 12.0 475.0	<input checked="" type="checkbox"/>		
Decision: 16	2.0 7.0 32.0 18.0 828.0	<input checked="" type="checkbox"/>		
Decision: 17	3.0 21.0 19.0 15.0 465.0	<input checked="" type="checkbox"/>		
Decision: 18	4.0 24.0 17.0 12.0 445.0	<input checked="" type="checkbox"/>		
Decision: 19	3.0 22.0 18.0 15.0 461.0	<input checked="" type="checkbox"/>		
Decision: 20	4.0 16.0 25.0 12.0 575.0	<input checked="" type="checkbox"/>		
Decision: 21	2.0 17.0 22.0 20.0 484.0	<input checked="" type="checkbox"/>		
Decision: 22	2.0 5.0 34.0 24.0 756.0	<input checked="" type="checkbox"/>		
Decision: 23	3.0 17.0 23.0 15.0 495.0	<input checked="" type="checkbox"/>		

Fig. 6. Solution resultset sample enables the user to select preferred solutions

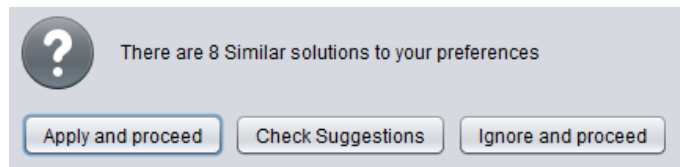


Fig. 7. A sample of Naïve Bayes suggestions to the user

Each result is displayed in a window detailing solution's objective values and solution features details.

III. ALGORITHMS USED, HOW AND WHY?

Smart OptiSelect uses hybrid of Multiobjective optimization and machine learning algorithms to achieve effective User-In-the-Loop preference based framework.

A. IBEA

Indicator-based evolutionary algorithm (IBEA) is a multi-objective evolutionary algorithm that can be combined with arbitrary indicators. In contrast to existing algorithms, IBEA can be adapted to the preferences of the user and, moreover does not require any additional diversity preservation mechanism; such as fitness sharing to be used. IBEA calculates domination value (i.e. amount of dominance) based on indicator (e.g. hypervolume). It favors objectives, i.e. user preferences.

A comparison among various multi-objective search-based software engineering methods was performed by A. Sayyad et al. [10]. It has shown that IBEA performs much better in product line feature optimization than methods in widespread use especially with increased number of optimization objectives. IBEA works best since it makes most use of user preference knowledge. It also generates far more products without violations of domain constraints.

To adopt IBEA, jMetal [11] IBEA library was used by Smart OptiSelect through formatting feature model trees attributes into indexed formats ready for the evaluation process. Then, it passes problem to IBEA in a binary-encoded-problem format. IBEA generates selected/non-selected features list for each decision based on the optimization of features selection using hyper volume indicator.

B. NSGAI

NSGA-II [12] is a multi-objective evolutionary algorithm which uses a non-dominated sorting for optimizing multi-objective problems. It is able to find high spread solutions in all problems. It pays special attention towards creating a diverse Pareto-optimal front within low computational requirements, elitist approach, and parameter-less sharing approach.

NSGA-II Calculates distance to the closest point for each objective. The fitness is the product of these distances. It favors higher fitness, i.e. more isolated points. It favors absolute domination and more spread out solutions.

NSGA-II came second after IBEA in optimizing product lines feature models [10] achieving better spread and hyper volume rather than rest of multi-objective evolutionary algorithms.

JMetal [11] NSGA-II library was used by Smart OptiSelect as following:

- Feature model attributes tree is reformatted into an indexed array to speed up evaluation processes.
- Problem is passed to NSGA-II as a binary-encoded problem using selected/non-selected features for each decision.
- NSGA-II generates optimized solution set based on maximizing the spread of features attributes.

C. Hybrid Optimization

Smart OptiSelect runs both optimization algorithms independently for a fixed amount of time rather than fixed amount of evaluations to control the performance and to guarantee each of optimization algorithms is not waiting for other. Then both algorithms solutions are merged, ranked and filtered.

For achieving this merging process, employing Pareto front ranking [13] gave a way to extract non-dominated solutions with highest ranks from multiobjective optimization hybrid solutions.

After each phase of the optimization process, solutions generated by both algorithms are plotted on the fitness space as shown in Fig. 8. J Metal Library [14] is used to sort, filter and extract first rank of Pareto front optimum solutions.

D. Naïve Bayes

Naïve Bayes [15] classifier is selected for providing suggestions to the user based on his preferred solutions selected from totals solutions result set.

The Naive Bayes algorithm is a simple probabilistic classifier that calculates a set of probabilities by counting the frequency and combinations of values in a given data set.

The probability of a specific feature in the data appears as a member in the set of probabilities derived by calculating the frequency of each feature value within a class of a training data

set. The training dataset used to train a classifier algorithm by using known values to predict future, unknown values.

Although Naïve Bayes performed consistently worse than C4.5 [16], it remained true to its reputation and sufficient enough for being used for providing suggestions to the user for following reasons:

- Its probabilistic nature depending on counting frequency and combination given in training set suited well the problem in hand as training dataset is the same of test dataset.
- It can build models from extremely small feature sets [17].
- Its simplicity and fairly competitive performance make it the best alternative.

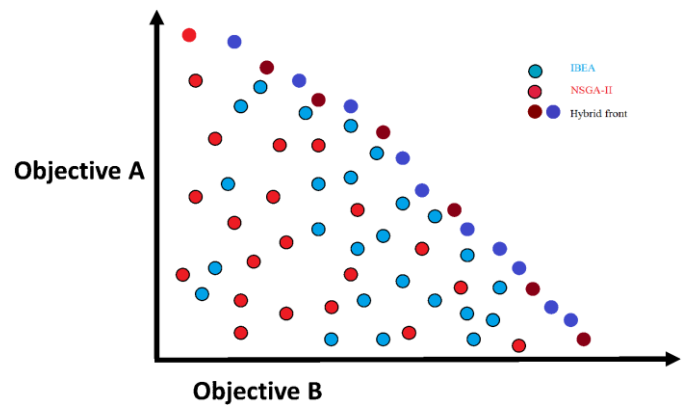


Fig. 8. Hybrid ranked non-dominant optimization solutions.

Smart OptiSelect used Naïve Bayes through following implementation: Given a set of r decision vectors $D = \{d_1, \dots, d_r\}$, classified along a two C classes, $C = \{c_1, c_2\}$ for representing Selected/Non-Selected classes, Bayesian classifiers estimate the probabilities of each class c_k given a decision d_j as:

$$P(c_k|d_j) = (P(c_k)P(d_j | c_k)) / P(d_j) \quad (1)$$

In eq. 1, $P(d_j)$ is the probability that a randomly picked decision has vector d_j as its representation, and $P(c_k)$ the probability that a randomly picked decision belongs to c_k .

$P(d_j | c_k)$ is the product of the probabilities of each feature that appears in the decision. So, $P(d_j | c_k)$ may be estimated as:

$$P(d_j | c_k) = \prod_{i=1}^{|T|} P(F_{ij} | c_k) \quad (2)$$

Where, $d_j = (f_1, \dots, f_{|T|})$.

For classifying datasets, Weka library implementation was adopted by OptiSelect. Weka is a data mining library contains many machine learning algorithms [18].

Smart OptiSelect uses Weka Naïve Bayes library through the following steps:

- 1) For each decision, the application checks each feature in it and format it into binary map represents presence and absence of that feature.
- 2) Training Naïve Bayes using every decision and its corresponding category (Selected/Non-Selected).
- 3) While testing a decision, algorithm calculates the probability of each feature of the test decision.
- 4) The test decision is classified into Selected/Non-Selected categories on the basis of probability.

E. C4.5

C4.5 [19] is adopted by Smart OptiSelect to build user preferences decision tree based on user’s preferred solutions. This decision tree evolves along optimization increments and is used to determine the user preferences. During each framework cycle, the results from the optimization process are filtered using the C4.5 built preference decision tree during previous cycles.

C4.5 may perform slightly worse than Support Vector Machine and Random Forest algorithms in terms of output quality, yet it is the most convenient to be used by Smart OptiSelect for its superiority in building models from extremely small feature sets [17].

C4.5 is based on inductive logic programming methods, constructing a decision tree based on a training set of data and using an entropy measure to determine which features of the training cases are important to populate the leaves of the tree.

The algorithm first identifies the dominant attribute of the training set and sets it as the root of the tree. Second, it creates a leaf for each of the possible values the root can take. Then, for each of the leaves it repeats the process using the training set data classified by this leaf. The core function of the algorithm is determining the most appropriate attribute to best partition the data into various classes.

Smart OptiSelect uses C4.5 through the following steps:

- 1) After each iteration, C4.5 is trained to build decision tree using user selected preferred decisions as a training set using two classes (Selected/Non-Selected).
- 2) After finishing each next optimization cycle, each decision is tested using the C4.5 built decision tree to calculate decisions belonging to the user’s preferences class, resulting in a filtered solution result set.

F. Mechanism Design Methodologies

Feature management conflict control: During the phase of feature attributes’ management, the user is allowed to configure forcing and excluding specific features. This type of management may violate feature model mandatory constraint or cross tree constraints.

The pseudo code shown in Fig. 9 illustrates how the application deals with such probable conflicts.

Pre-optimization indexing: Performance and memory management are essential especially when searching large feature model trees attached with dynamic multi-objective attributes. A sorted index array is introduced to hold references for tree features nodes as shown in Fig. 10.

```

IF control is exclusion/forcing THEN
  Get all successors/parents affected nodes
  FOR each _node in affected nodes
    If _node exclusion/forcing causes conflict
      THEN
        Get all corrective alternatives
        If corrective alternative count > 1 THEN
          Notify the user
        ELSE
          Perform corrective action
        END IF
      END IF
    END FOR
  END FOR

```

Fig. 9. Feature management consequences control pseudo code

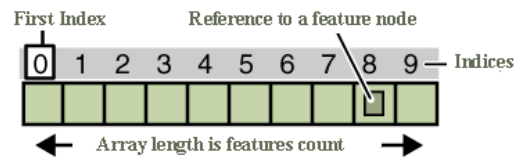


Fig. 10. Features Indexed List

Features tree is traversed using depth first algorithm once prior to optimization iteration to generate a sorted index array. This Index provides O (1) direct access to features properties and attributes. Hence, evaluation processes and search performance are optimized by avoiding tree repetitive search and tree diving recursion overhead which costs O (N). Thus, attributes are demoralized to a binary array.

Tree mutation probability: Based on the knowledge of SPLOT feature model tree structure, Tree mutation using special tree mutation probability parameter is used [20]. It aims to prevent mutations which violates feature model constraints and performs mutations with paying respect to feature model tree structure and constraints as shown in Fig. 11.

Usually, in the experiments, we set tree mutation probability to 0 to prevent tree structure and constraints violation while mutation. We also experimented raising the tree mutation probability parameter to %20 which resulted in more diversity in results but less correct solutions due to correctness thresholds.

```

FOR each bit in the decision string
  IF rand (0, 1) < mutation_probability THEN
    IF Deselecting root feature OR Deselecting a mandatory child
    feature whose parent is selected, or Group cardinality is violated AND
    rand (0,1) < tree_mutation_probability
    THEN
      Do not mutate
    ELSE
      Flip this bit
      IF selecting (turning on) a feature THEN
        Turn on children (a minimum skeleton)
      Else IF deselecting (turning off) a feature THEN
        Turn off all children
      END IF
    END IF
  END IF
END IF
END FOR
    
```

Fig. 11. Tree mutation procedure pseudo code

IV. USERS SATISFACTION RESULTS

Smart OptiSelect can display four result sets formats of solutions:

- IBEA optimization Result
- NSGAII optimization Result
- Hybrid Pareto front optimization result
- C4.5 preferences filtered result

A survey has been created among 20 specialized software project managers, software architects and system analysts. Each of them has run through the framework for five iterations then was asked to express his satisfaction with different versions of output. User satisfaction is expressed in terms of solutions richness and its relevance to the scope. Each user was only allowed to select one result set as the best result set based on his satisfaction for each iteration round. We calculated average satisfaction for each number of iterations round. Sometimes, one result version achieved much higher satisfaction than others. Other times, more than one result were nearly equaled as shown in Table 1.

TABLE I. SYSTEM OUTPUTS USERS SATISFACTION

	0 Iteration	1 Iteration	2 Iterations	3+ Iterations
IBEA Result	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
NSGAII Result	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Hybrid Result	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
C4.5 Result	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

The results have shown that:

IBEA results we generally more satisfying than NSGAII results because they made more attentions to users’ objectives.

Hybrid results attracted attention as it displayed interesting decision solutions added from NSGA-II.

During the first iterations, Users were more satisfied with IBEA and hybrid results as they have more decisions displayed than filtered result sets by C4.5.

Starting from second iteration, most of users - who paid an interest in certain solutions’ features - found that the C4.5 results were more convenient to their needs.

V. RELATED WORK DISCUSSIONS AND COMPARISON

Botterweck G. [21] feature configuration tool S2T2 Configurator integrates a visual interactive representation of the feature model and a formal reasoning engine that calculates consequences of the user’s actions and provides formal explanations. Still it didn’t provide a multi-objective support nor incremental configuration.

FAMA [22] is a framework for the automated analysis of feature models integrating some of the most commonly used logic representations and solvers proposed for automated analyses of feature models.

The Feature Model Plugin (FMP) [23] is implemented as an Eclipse plug-in. It supports configuration based on feature diagrams. But it does not have the analysis of FMs among its main goals. It does not support attributed feature models.

CaptainFeature is a feature modelling tool using the FODA notation to render and configure feature diagrams. It does not support the automated analysis of FMs.

\ [24] is a lightweight yet expressive language for structural modeling: feature modeling and configuration, class and object modeling.

TABLE II. SUMMARY OF FEATURE CONFIGURATION PROPOSAL

	Feature Model Representation	Iterative Configuration	Multi-Objective Optimization	Hybrid Optimization	Features Automated Analysis	Attributed Feature Model	Machine Learning Based Preferences
S2T2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
FAMA	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
FMP	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
CaptainFeature	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Clafer	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>

Opti-Select	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Smart OptiSelect	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

VI. CONCLUSION AND FUTUTRE WORK

Smart OptiSelect pays more attention to user preferences by recoding his selections and training the framework incrementally to narrow the results around selected decisions and solutions.

Smart OptiSelect is considered an innovative framework as it is the first in the field of product-lines-search-based-optimization to adopt and purpose the following techniques and algorithms, as well as merging their outputs together consistently in one frame work application:

- Incremental optimization: The user can run feature-selection optimization process in increments allowing the user to adjust both the objectives and attributes in the middle of the optimization process, and to set preferred solutions.
- Hybrid Optimization: The Innovative technique utilizing the superiority of IBEA and NSGA-II [25] [26] in the field of search-based-product-line-optimization, as well as merging and filtering their results using Pareto front ranking.
- Utilization of machine learning techniques such as Naïve Bayes and C4.5 for their capability to build classifiers and decision trees to produce preference-based-solutions inspired by the user’s selections among optimization increments.

Through our continuous research and development, our future steps will be:

- Using machine learning techniques to train classifiers to learn the user’s objectives classification and categorization. This may vary as a simple objective or a certain relation between some features rather than his preferred features.
- Utilization of newly proposed 10-WS-C4.5-TDM-NB-TDMR [27] for user’s preferences classification problem.
- Examining scalability of the results obtained with larger feature models, such as the Linux kernel feature model (part of LVAT repository [28]) composed of 6888 features.

ACKNOWLEDGMENT

Our thanks to Dr. Abdel Salam Sayyad, Dr. Tim Menzis and to Dr. Hany Ammar from West Virginia University for their valuable advices and providing access to benchmarks.

REFERENCES

- [1] Yamany, El, Ahmed Eid, Mohamed Shaheen, and Abdel Salam Sayyad. "OPTI-SELECT: an interactive tool for user-in-the-loop feature selection in software product lines." In Proceedings of the 18th International Software Product Line Conference: Companion Volume for Workshops, Demonstrations and Tools-Volume 2, pp. 126-129. ACM, 2014.
- [2] Sayyad, Abdel Salam, Tim Menzies, and Hany Ammar. "On the value of user preferences in search-based software engineering: A case study in software product lines." In Software Engineering (ICSE), 2013 35th International Conference on, pp. 492-501. IEEE, 2013.
- [3] Mendonca, Marcilio, Moises Branco, and Donald Cowan. "SPLOT: software product lines online tools." In Proceedings of the 24th ACM SIGPLAN conference companion on Object oriented programming systems languages and applications, pp. 761-762. ACM, 2009.
- [4] Sayyad, Abdel Salam, Joseph Ingram, Tim Menzies, and Hany Ammar. "Optimum feature selection in software product lines: Let your model and values guide your search." In Combining Modelling and Search-Based Software Engineering (CMSBSE), 2013 1st International Workshop on, pp. 22-27. IEEE, 2013.
- [5] Zitzler, Eckart, and Simon Künzli. "Indicator-based selection in multiobjective search." In Parallel Problem Solving from Nature-PPSN VIII, pp. 832-842. Springer Berlin Heidelberg, 2004.
- [6] Sadeghi, Javad, Saeid Sadeghi, and Seyed Taghi Akhavan Niaki. "A hybrid vendor managed inventory and redundancy allocation optimization problem in supply chain management: An NSGA-II with tuned parameters." Computers & Operations Research 41 (2014): 53-64.
- [7] Kumar, Rajeev, and Peter Rockett. "Improved sampling of the Pareto-front in multiobjective genetic optimizations by steady-state evolution: a Pareto converging genetic algorithm." Evolutionary computation 10, no. 3 (2002): 283-314.
- [8] Ting, S. L., W. H. Ip, and Albert HC Tsang. "Is Naive Bayes a good classifier for document classification?." International Journal of Software Engineering and Its Applications 5, no. 3 (2011): 37.
- [9] Quinlan, J. Ross. "Improved use of continuous attributes in C4.5." arXiv preprint cs/9603103 (1996).
- [10] Sayyad, Abdel Salam. "Evolutionary Search Techniques with Strong Heuristics for Multi-Objective Feature Selection in Software Product Lines." PhD diss., WEST VIRGINIA UNIVERSITY, 2014.
- [11] Nebro, Antonio J., and Juan J. Durillo. "jMetal 4.5 User Manual." (2014).
- [12] Deb, Kalyanmoy, Samir Agrawal, Amrit Pratap, and Tanaka Meyarivan. "A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II." Lecture notes in computer science 1917 (2000): 849-858.
- [13] Bosman, Peter AN. "On gradients and hybrid evolutionary algorithms for real-valued multiobjective optimization." Evolutionary Computation, IEEE Transactions on 16, no. 1 (2012): 51-69.
- [14] Matjelo, Naleli Jubert, Fred Nicolls, and Neil Muller. "Evaluation of Optimal Control-based Deformable Registration Model." In New Trends in Networking, Computing, E-learning, Systems Sciences, and Engineering, pp. 117-124. Springer International Publishing, 2015.
- [15] Zhang, Harry. "The optimality of naive Bayes." AA 1, no. 2 (2004): 3.
- [16] Dimitoglou, George, James A. Adams, and Carol M. Jim. "Comparison of the C4.5 and a Naive Bayes Classifier for the Prediction of Lung Cancer Survivability." arXiv preprint arXiv:1206.1121 (2012).
- [17] Vatolkin, Igor, Mike Preuß, and Günter Rudolph. "Multi-objective feature selection in music genre and style recognition tasks." In Proceedings of the 13th annual conference on Genetic and evolutionary computation, pp. 411-418. ACM, 2011.
- [18] Sharma, Narendra, Aman Bajpai, and Mr Ratnesh Litoriya. "Comparison the various clustering algorithms of weka tools." facilities 4 (2012): 7.
- [19] Ruggieri, Salvatore. "Efficient C4.5 [classification algorithm]." Knowledge and Data Engineering, IEEE Transactions on 14, no. 2 (2002): 438-444.

- [20] Linsbauer, Lukas, Roberto Erick Lopez-Herrejon, and Alexander Egyed. "Feature Model Synthesis with Genetic Programming." In *Search-Based Software Engineering*, pp. 153-167. Springer International Publishing, 2014.
- [21] Botterweck, Goetz, Mikolas Janota, and Denny Schneeweiss. "A Design of a Configurable Feature Model Configurator." *VaMoS 29* (2009): 165-168.
- [22] Benavides, David, Sergio Segura, Pablo Trinidad, and Antonio Ruiz Cortés. "FAMA: Tooling a Framework for the Automated Analysis of Feature Models." *VaMoS 2007* (2007): 01.
- [23] Czarnecki, Krzysztof, and Chang Hwan Peter Kim. "Cardinality-based feature modeling and constraints: A progress report." In *International Workshop on Software Factories*, pp. 16-20. 2005.
- [24] Antkiewicz, Michał, Kacper Bąk, Alexandr Murashkin, Rafael Olacchia, Jia Hui Jimmy Liang, and Krzysztof Czarnecki. "Clafar tools for product line engineering." In *Proceedings of the 17th International Software Product Line Conference co-located workshops*, pp. 130-135. ACM, 2013.
- [25] Peddabachigari, Sandhya, Ajith Abraham, Crina Grosan, and Johnson Thomas. "Modeling intrusion detection system using hybrid intelligent systems." *Journal of network and computer applications* 30, no. 1 (2007): 114-132.
- [26] Purshouse, Robin C., Kalyanmoy Deb, Maszatul M. Mansor, Sanaz Mostaghim, and Rui Wang. "A review of hybrid evolutionary multiple criteria decision making methods." *COIN Report*,(2014005), January (2014).
- [27] Molano, Viviana, Carlos Cobos, Martha Mendoza, Enrique Herrera-Viedma, and Milos Manic. "Feature Selection Based on Sampling and C4.5 Algorithm to Improve the Quality of Text Classification Using Naïve Bayes." In *Human-Inspired Computing and Its Applications*, pp. 80-91. Springer International Publishing, 2014.
- [28] She, Steven, Rafael Lotufo, Thorsten Berger, Andrzej Wasowski, and Krzysztof Czarnecki. "The Variability Model of The Linux Kernel." *VaMoS 10* (2010): 45-51.

Decision Support System for Learning Disabilities Children in Detecting Visual-Auditory-Kinesthetic Learning Style

Wan Fatin Fatimah Yahya, Noor Maizura Mohamad Noor

School of Informatics and Applied Mathematics

Universiti Malaysia Terengganu

Kuala Terengganu, Terengganu Malaysia

wanfatinyahya@gmail.com

Abstract—The innovation of information and communications technology in education has improved the learning quality and has provided a positive impact on the learning environment and its community. Integrating learning styles in adaptive e-Learning systems has been considered a growing trend in technology to improve the learning process. Also, when these technologies are obtainable, reasonable and available, they represent more than a transformation for people with disabilities. The purpose of this research is to adopt a decision support system in e-learning in order to model the visual-auditory-kinesthetic learning style focusing on learning disabilities children. Learning disabilities children face difficulties in processing and retaining information and thus have problems in the classroom. Providing adaptively based on learning styles has potential to make learning easier for students and increase learning progress. The traditional way to identify learning styles is by using questionnaires. Even though, the problem with the traditional approach is not all the students are interested to fill out a questionnaire. Hence the most recent years, several approaches have been proposed for automatically detecting learning styles to solve these problems. Therefore, the main aim of this paper is to propose e-learning decision support system architecture to estimate students' learning style automatically using literature-based method. Calculation to estimate each of the student's learning styles based on number of visits and the time that spent on learning objects with respect to the visual-auditory-kinesthetic learning style.

Keywords—*learning disabilities; decision support system; visual-auditory-kinesthetic learning style*

I. INTRODUCTION

Today Information and Communication Technologies (ICT) have been broadly applied to the field of education and learning technologies transformed educational systems with impressive progress. The enhanced use of ICT in most sectors of the community, especially in supporting education and inclusion for persons with disabilities can be a powerful tool to improve their quality of life. Many children with disabilities are facing a wide range of barriers, including omitted from educational opportunities and do not complete primary education [1]. The effective application of technologies can ensure comprehensive classroom learning, accessibility, teaching and learning content and techniques more in friendly with learners' needs. E-Learning appears as a new education paradigm to fulfill that learners need, overcome physical deficiency of the users and decrease barriers in education [1, 2, 3]. Some researchers have explored that accommodating learning styles-based approach in e-learning has proven to be effective in the classroom and allows individual learning styles and preferences to be accommodated [3, 4, 5]. Students with learning disabilities may have different learning difficulties from each other [3] and have different ways in their own

learning processes to help them learn better. These several well-known learning style models such as Kolb, Honey & Mumford, Dunn & Dunn and Felder-Silverman. To implement the adaptation in such e-Learning systems, students' learning styles need to be identified first.

There are many ways of identifying learners' learning style and commonly it is static approach that uses a questionnaire. This approach is still useful until now. But, the problem with the questionnaire approach is not all the students are interested to fill out the questionnaire and the result depends heavily on students' mood [6]. Moreover, these questionnaires are unable to detect changes in a learner's learning style [7]. As a result of these problems, various researches have been concentrated on how to identify learners' learning styles automatically. There are two broadly approaches, including data-driven approach (DDA) and literature-based approach (LBA). Generally, the idea of the automatic detection learning style can be simplified as Fig. 1. In our study, we choose the literature-based approach to automatically detect learning styles of learners in DSS e-Learning systems. In this research, we propose approach decision support systems for specific groups of users, by taking into consideration user's learning style in e-learning.

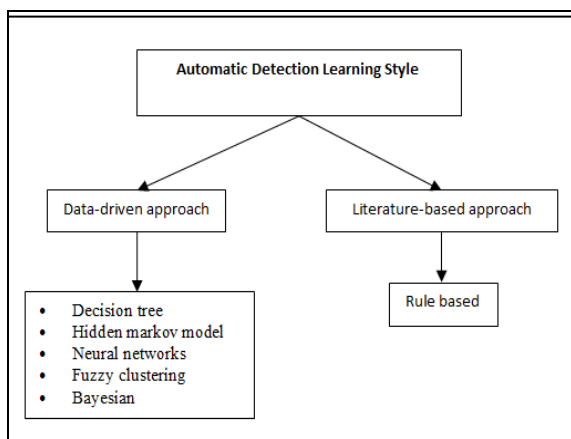


Fig. 1. Idea of automatic detection learning style

II. DECISION SUPPORT SYSTEM

A decision support system (DSS) is an interactive computer-based system capable of supporting decision-making process. Lately, DSS technology has been positively applied to many decision-making problems in numerous disciplines, including education [8]. DSS as part of e-learning systems can analyze data in users' profiles and allow the learners to select optimized learning paths based on previous learning information about learners [9]. Individuals with learning disabilities possibly will have dissimilar problems from each other. Therefore, it is supposed that learning surroundings which are developed for learning disability individual should be adapted to the learning need of the individual. DSS is suitable and has many advantages for education of learning disability students as they can deliver a different presentation of learning contents and recommends adaptive learning paths based on analyses previous learning activities [3, 8]. The DSS in our approach systems can be described as the systems that determine the students' learning styles automatically and delivers a different presentation of learning content for learner's with different learning styles.

III. LEARNING DISABILITIES CHILDREN

The learning disability is an umbrella term that describes of learning problems includes dyscalculia, dysgraphia, dyspraxia, central auditory processing disorder, non-verbal learning disorder, visual-spatial disorder, visual motor disorder, developmental aphasia and language disorders [10] such figured in Fig. 2.

The Malaysian Ministry of Education categorized learning disabilities students under special needs [11]. The Ministry of Education offers special education programs for hearing, visual and learning disabilities students [12]. Learning disabilities child differ from each other and may not have the same learning problems as another child with LD. Most of these children face a wide range of problems in educational chances

and not completed primary education [1]. There is no treatment for learning disabilities, but they can be high achievers and learn successfully with the right help [1, 13].

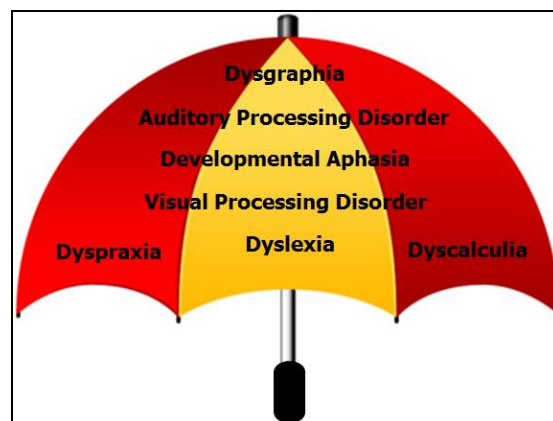


Fig. 2. Learning disabilities

Numerous researches have shown that emerging learning environments blended with technology can play an important role in specific disadvantaged groups such as the blind, those with movement disabilities and LD. It also gives a great potential to support and enhance students learning processes to live freely and learn easier [3, 14, 15]. It is a good transformation and opportunity for the LD people solve the problems that happen in traditional educational systems. Traditional computer learning environments proposes the same content and they do not consider the individual differences, preferences and interests [3].

IV. LEARNING STYLE

According to Keefe, define learning styles as 'cognitive, affective, and physiological traits that serve as relatively stable indicators of how learners perceive, interact with, and respond to learning environments'. By identifying student's learning style, teachers should be encouraged and provide to create a learning process sensitive with the students' learning needs [16, 17]. Learning styles are significant in the learning process since they may help student's achievement is improved and would be to increase self-awareness of the strengths and weaknesses [18, 19, 20].

There are numerous models of learning styles from the literature, like Felder- Silverman, Dunn and Dunn, Honey and Mumford, Kolb and Visual- Auditory-Kinesthetic (VAK) learning style model [21]. VAK and Felder are two well-known models used in adaptive e-learning system [20]. The main purpose for selecting a VAK learning style of our work that this learning style is most widely-used, simple, suitable for children and identify a student's dominant mode of perceiving information [22, 23, 24]. Fig.3 shows the features VAK learning style model.



Fig. 3. VAK learning style

The VAK learning style model focuses on human observation channel vision, hearing and feeling. This model is categorizations into three modalities, firstly visual learners, secondly auditory learners and lastly kinesthetic learners or tactile learners [23, 25, 26]. Visual learners prefer to learn via seeing. For these learners, pictures, flow diagrams and videos are the best learning instruments. Auditory learners have a preference for listening, audibly and learn best by hearing. Kinesthetic learners' best learn through feeling or doing-experiencing such as moving, touching, and doing [27]. For these learners, computer games, interactive animations are the best learning instruments [21, 23]. Based on each mode's tendency, automatic learning style detection is conducted to obtain students' feedback on computer based learning.

V. METHOD

Because of the limitations and disadvantages of questionnaire-based learning style detection, the detection process must be computerized. From previous studies, it is clear that the process of automatic detection of learning styles includes fundamentally two stages. The first stage is identifying the significant behaviour for each learning style and the second stage is inferring the learning style from the behavior and actions of an individual [7, 28] (see Fig. 4). Identifying the significant behaviour for each learning style involves three phases, firstly is choosing the relevant features of behaviour, secondly is categorizing the occurrence of the behaviour and last one is defining the patterns for each element of the learning style. For the identifying the significant behaviour for each learning style involves of the three phases, firstly is choosing the relevant features of behaviour, secondly is categorizing the occurrence of the behaviour and last one is defining the patterns for each element of the learning style. Then the calculation methodology can be data-driven or literature-based approach in the inferring learning style stage [29].

A. Literature-Based

The literature-based approach is to use the behaviour of students in order to get suggestions about their learning style

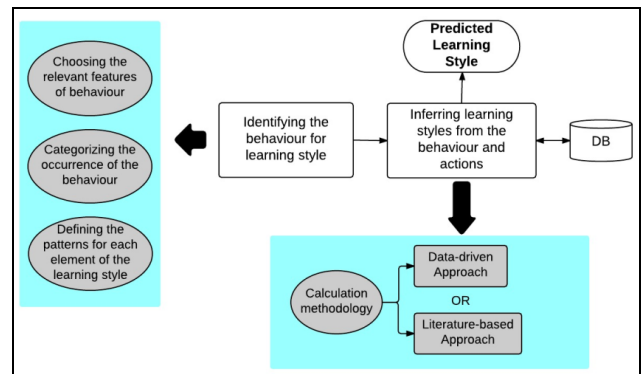


Fig. 1 Concept of automatic detection of learning style

preferences [16]. This approach was proposed by Graf et al. [30]. Then a simple rule-based method is applied to calculate learning styles from the number of matching hints. This approach is same to the technique used for calculating learning styles in the Index of Learning Styles (ILS) questionnaire and has the benefit to be nonspecific and relevant for data assembled from any course [31].

Several studies have been used in literature-based approach, such as, Graf, et al. [30], which first proposed the new methodology of literature-based approach for automatic detection of styles preferences according to the Felder-Silverman learning style model (FSLSM), in Learning Management Systems (LMS). Simsek, et al. [32] recommended a literature-based approach for automatic student modelling taking into consideration the learner interface interactions. George, et al. [7] propose to use a mix of data-driven approach and literature-based approach. Dung and Florea [33] use literature-based approach for automatic detection of learning style and use the number of visits and time that the learner spends on learning objects as parameters. Ahmad, et. al [16] use literature-based approach to analyzed pattern of behaviour for Malaysian polytechnic students who studied Interactive Multimedia course. In our approach, we use the VAK learning style model and follow literature-based approach and used a simple rule engine to estimate the learning style.

B. Learning Styles Estimation

Literature-based method is used to estimate learning styles automatically. Calculation to estimate each of the student's learning styles based on number of visits and the time that are spent on learning objects. Learning object can be defined as "any digital resource that can be reused to support learning" [31]. This meaning contains all that can be delivered across the network on request. Examples of digital resources include digital images or photos, live or prerecorded video or audio snippets, small bits of text, animations, smaller web-delivered applications, multiple choice exercise, and book [34].

The predictable time spent on each learning object, $Time_{predictible}$, is determined. The time that a learner actually spent on each learning object, $Time_{spent}$, is recorded. For example, if $Time_{predictible}$ of a visual learning object is 30 second. After a period of time X, sums of $Time_{spent}$ for three learning style elements of the learner is calculated. Then, the ratios of time (RT) are found out as the formula in eq.1.

$$RT_{LS} = \frac{\sum Time_{spent}}{\sum Time_{predictable}} \quad (1)$$

To calculate the ratio of number_visit, $RV_{LS_element}$, number of learning object visited, Ex_{visit} , and total of learning object, Ex each learning style element are compute using the formula as eq. 2.

$$RV_{LS_element} = \frac{\sum Ex_{visit}}{\sum Ex} \quad (2)$$

Finally, the average ratios, R_{avg} , are calculated as the formula in eq. 3.

$$R_{avg} = \frac{(RT + RV)}{2} \quad (3)$$

Then learning style is estimated based on the simple rule as shown in Table 1:

TABLE I. SIMPLE RULE ESTIMATION OF LEARNING STYLES

Ravg	LS
0 – 0.3	Weak
0.3 – 0.7	Moderate
0.7 – 1	Strong

VI. RESULTS AND DISCUSSIONS

In order to meet individual user’s needs to teaching delivery, LS models have to be established and integrated within e-learning. The review as showed in this study the significant role that reliable learning style models can play in enhancing the learning ability and learning background if these are well-matched. Numerous past studies have shown that the VAK learning style model which represents one of the commonly used, very simple and suitable for children [22, 23, 24 ,27]. Previous study presented that automatic approach as a better approach to identify learning style in online learning because it is based on the actual students’ behaviour pattern while learning [35]. The purpose of this research, literature-based approach is chosen. This approach use rule based method for identifying learning style. According to Graf [29], the main strength of literature-based approach is the ability of deducing LS without needing training data and depends directly on learning style model. We construct our own research architecture DSS e-learning for LD children and to be used conveniently in the future.

A. Research Architecture

In this section, we comprehensively describe the architecture of DSS e-learning for LD children and demonstrate the individual components needed to implement our approach. The proposed architecture is represented in Fig.

6. The structure of the system consists of user interface, adaptive engine, content service, system manager and personal profile service. The user interface deals with the learners’ registration and this service is responsible for user login. The service will add information in profile model database and communicate with the learners’ personal profile service agent, The Adaptive Engine (AE), is responsible for suggesting the learning styles according to learners behaviours. Content Service control what type of learning materials should be provided to each learner based on their learning styles. System Manager allows the teachers to update the content. The personal profile service saves the users’ information from the profile model database and communicate with the adaptive engine in order to decide to deliver the right learning materials

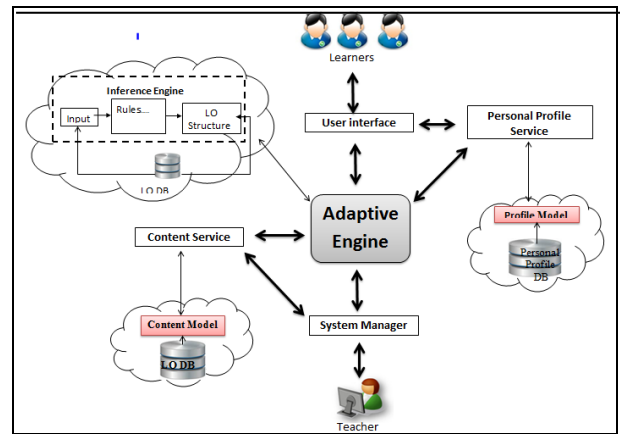


Fig. 2. The architecture of the System

VII. CONCLUSION

The main conclusion of the present study is that the future research should take into account that LD students need to be encouraged to learn something they like. Based on that, we will design a DSS e-learning system that adapts to learners’ preference of learning style automatically. We propose an e-learning DSS system architecture to detect the students’ learning styles automatically using a literature-based method. This method is based on the rule base for calculating to estimate each of the student’s learning styles based on number of visits and the time that he or she spent on learning objects. Our future work will concentrate on extensive development in validating the proposed system and the efficiency of the method.

ACKNOWLEDGMENT

We would like to thank we would like to thank vote 63920 of research and development in RFID technologies from SENSTECH Sdn. Bhd., Malaysia for contributing funds to conduct this research. the authors would also like to express our deepest gratitude to the anonymous reviewers of this paper. their useful comments have played a significant role in improving the quality of this work.

REFERENCES

- [1] M. Laabidi, M. Jemni, L. Jemni Ben Ayed, H. Ben Brahim and A. Ben Jemaa, "Learning technologies for people with disabilities," *Journal of King Saud University - Computer and Information Sciences*, vol. 26, pp. 29-45, 2014.
- [2] S. Abu-Naser, A. Al-Masri, Y. A. Sultan and I. Zaqout, "A prototype decision support system for optimizing the effectiveness of elearning in educational institutions," *International Journal of Data Mining & Knowledge Management Process(IJDKP)*, vol. 1, pp. 1-13, 2011.
- [3] E. Polat, T. Adiguzel and O. E. Akgun, "Adaptive Web-Assisted Learning System for Students with Specific Learning Disabilities: A Needs Analysis Study," *Educational Sciences: Theory and Practice*, vol. 12, pp. 3243-3258, 2012.
- [4] H. Ben Brahim, A. Ben Jemaa, M. Jemni and M. Laabidi, "Towards the Design of Personalised Accessible E-Learning Environments," in *Advanced Learning Technologies (ICALT), 2013 IEEE 13th International Conference 2013*, pp. 419-420.
- [5] M. Laabidi and M. Jemni, "Personalizing Accessibility to E-Learning Environments," in *Advanced Learning Technologies (ICALT), 2010 IEEE 10th International Conference 2010*, pp. 712-713.
- [6] Q. D. Pham and A. M. Florea, "A method for detection of learning styles in learning management systems," *UPB Scientific Bulletin, Series C: Electrical Engineering*, vol. 75, pp. 3-12, 2013.
- [7] George Abraham, Balasubramanian V. and R. K. Saravanaguru, "Adaptive e-Learning Environment using Learning Style Recognition," *International Journal of Evaluation and Research in Education (IJERE)*, vol. 2, pp. 23-31, 2013.
- [8] A. A. Kardan and H. Sadeghi, "A Decision Support System for Course Offering in Online Higher Education Institutes," *International Journal of Computational Intelligence Systems*, vol. 6, pp. 928-942, 2013/09/01 2013.
- [9] M. Yarandi, H. Jahankhani and A.-R. H. Tawil, "Towards Adaptive E-Learning using Decision Support Systems," *International Journal of Emerging Technologies in Learning*, 2013.
- [10] P. Shajimon and S. S. P. Jose, "Strategy for the physical and social development of learning disabled," *Learning Disability*, p. 230.
- [11] T. HJ, C. SK and W. PJ, "Student Learning Disability Experiences, Training And Services Needs Of Secondary School Teachers," *Malaysian Journal of Psychiatry*, 2010.
- [12] M. M. Ali, R. Mustapha and Z. M. Jelas, "An Empirical Study on Teachers' Perceptions towards Inclusive Education in Malaysia," *International Journal of Special Education*, vol. 21, pp. 36-44, 2006.
- [13] J. M. David and K. Balakrishnan, "Prediction of Learning Disabilities in School-Age Children using SVM and Decision Tree," *Int. J. of Computer Science and Information Technology*, ISSN, pp. 0975-9646, 2011.
- [14] N. A. Beacham and J. L. Alty, "An investigation into the effects that digital media can have on the learning outcomes of individuals who have dyslexia," *Computers & Education*, vol. 47, pp. 74-93, 2006.
- [15] T. Adam and A. Tatnall, "Using ICT to improve the education of students with learning disabilities," vol. 281, ed, 2008, pp. 63-70.
- [16] N. Ahmad, Z. Tasir, J. Kasim and H. Sahat, "Automatic Detection of Learning Styles in Learning Management Systems by Using Literature-based Method," *Procedia - Social and Behavioral Sciences*, vol. 103, pp. 181-189, 2013.
- [17] J. G. Sharp, R. Bowker and J. Byrne, "VAK or VAKuous? Towards the trivialisation of learning and the death of scholarship," *Research Papers in Education*, vol. 23, pp. 293-314, 2008.
- [18] J. Feldman, A. Monteserin and A. Amandi, "Automatic detection of learning styles: state of the art," *Artificial Intelligence Review*, pp. 1-30, 2014/05/15 2014.
- [19] R. Mokhtar, S. N. H. S. Abdullah and N. A. M. Zin, "Classifying modality learning styles based on Production-Fuzzy Rules," in *Pattern Analysis and Intelligent Robotics (ICPAIR), 2011 International Conference on*, 2011, pp. 154-159.
- [20] H. D. Surjono, "The Evaluation of a Moodle Based Adaptive e-Learning System," *International Journal of Information & Education Technology*, vol. 4, 2014.
- [21] F. A. Khan, E. R. Weippl and A. M. Tjoa, "Integrated Approach for The Detection of Learning Styles and Affective States," in *World Conference on Educational Multimedia, Hypermedia and Telecommunications*, 2009, pp. 753-761.
- [22] S. Gholami and M. S. Bagheri, "Relationship between VAK Learning Styles and Problem Solving Styles regarding Gender and Students' Fields of Study," *Journal of Language Teaching and Research*, vol. 4, pp. 700-706, 2013.
- [23] W. a. Qutechate, T. Almarabeh and R. Alfayez, "E-Learning System In The University Of Jordan: Problem Solving Case Study," *Journal of Theoretical & Applied Information Technology*, vol. 53, 2013.
- [24] U. Ocepek, Z. Bosnić, I. Nančovska Šerbec and J. Rugej, "Exploring the relation between learning style models and preferred multimedia types," *Computers & Education*, vol. 69, pp. 343-355, 2013.
- [25] C. Wolf, "iWeaver: towards' learning style'-based e-learning in computer science education," in *Proceedings of the fifth Australasian conference on Computing education-Volume 20*, 2003, pp. 273-279.
- [26] N. D. Fleming, "I'm different; not dumb. Modes of presentation (VARK) in the tertiary classroom," in *Research and Development in Higher Education*, *Proceedings of the 1995 Annual Conference of the Higher Education and Research Development Society of Australasia (HERDSA)*, HERDSA, 1995, pp. 308-313.
- [27] N. Othman and M. H. Amiruddin, "Different Perspectives of Learning Styles from VARK Model," *Procedia - Social and Behavioral Sciences*, vol. 7, pp. 652-660, 2010.
- [28] B. Velusamy and S. Anoucia Margret, "A narrative review of research on learning styles and cognitive strategies," *Journal of Theoretical and Applied Information Technology*, vol. 52, pp. 23-29, 2013.
- [29] S. Graf, "Adaptivity in learning management systems focussing on learning styles," *Faculty of Informatics, Vienna University of Technology*, 2007.
- [30] S. Graf, Kinshuk and L. Tzu-Chien, "Identifying Learning Styles in Learning Management Systems by Using Indications from Students' Behaviour," in *Advanced Learning Technologies, 2008. ICALT '08. Eighth IEEE International Conference on*, 2008, pp. 482-486.
- [31] P. Q. Dung and A. M. Florea, "A Literature-based Method to Automatically Detect Learning Styles in Learning Management Systems," in *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics*, 2012, p. 46.
- [32] Ö. Şimşek, N. Atman, M. M. İnceoğlu and Y. D. Arikan, "Diagnosis of Learning Styles Based on Active/Reflective Dimension of Felder and Silverman's Learning Style Model in a Learning Management System," in *Computational Science and Its Applications-ICCSA 2010*, ed: Springer, 2010, pp. 544-555.
- [33] P. Q. Dung and A. M. Florea, "An approach for detecting learning styles in learning management systems based on learners' behaviours," *International Proceedings of Economics Development & Research*, vol. 30, 2012.
- [34] T. Gaikwad and M. A. Potey, "Personalized Course Retrieval Using Literature Based Method in e-Learning System," in *Technology for Education (T4E), 2013 IEEE Fifth International Conference on*, 2013, pp. 147-150.
- [35] F. A. Dorça, L. V. Lima, M. A. Fernandes and C. R. Lopes, "Comparing strategies for modeling students learning styles through reinforcement learning in adaptive and intelligent educational systems: An experimental analysis," *Expert Systems with Applications*, vol. 40, pp. 2092-2101, 2013.

Identification of Potential Crime Tactical Path-Finding Using Analytical Hierarchy Process (AHP) in Situational Crime Prevention Crime Intelligence in New Era

Wan Mohd Farhan Bin Wan
Nawawi
Msc Computer Science
Universiti Malaysia Terengganu
Malaysia
GSK1830@pps.umt.edu.my

Noor Maizura Mohamad Nor
Lecturer Computer Science
Universiti Malaysia Terengganu
Malaysia
maizura@umt.edu.my

Masita Abdul Jalil
Lecturer Computer Science
Universiti Malaysia Terengganu
Malaysia
masita@umt.edu.my

Abstract— Most route guidance researches are mainly focused on route guidance for vehicles like Papago and Waze software. However, due to the recent spread of personal computing devices such as PDA, PMP and smart phone, route guidance for pedestrians is increasingly in demand especially in context of crime situation. The pedestrian route guidance is different from vehicle route guidance because pedestrians are affected more surrounding environment than vehicles. In this area of study which is situational crime prevention (SCP), pedestrian will try analyze the situation first before take any decision path. It satisfied goal of SCP in order to manipulate crime environment so that offender will seem harder and riskier, change criminal's ideas and reduce opportunities for criminal to commit crime. To solve this problem, we designed a model containing Multi Criteria Decision Making (MCDM) technique that is Analytical Hierarchy Process (AHP) to handle the uncertainty situations. Pedestrian path finding needs considerations of various factors affecting their safety walking. Factors affecting safety walking consists of 3 categories – distance between pedestrians and criminal, visibility of criminal view and obstacles frequency that exist in crime location. An application adopting the AHP idea was developed to calculate the weights of the criteria for evaluating each crime factors. The highest degree of AHP result will drive the pedestrians to choose best path-finding.

Keywords—personal computing devices, route guidance, situational crime prevention (SCP), multi criteria decision making (MCDM), analytical hierarchy process (AHP), crime situation.

I. INTRODUCTION

Since 2009s, the statistic of crime has been start decreased after 6th Prime Minister of Malaysia make a new approach by developing National Key Results Area (NKRA). Growth of crime technologies are distributed well like CCTV or expert report and CPTED approach already set up at hot hotspots and seem many reduction of crime statistic were happened. However, it facing one problem which is not yet solved by government which is feeling fear towards crime still high among citizens [1]. By realize this problem, our members tried to educate citizens which path are the most preferable to pass through since this alternatives provided the best method than others [2]. As we know, lack of awareness towards crime also leads high percentage to crime itself [3]. This effort will help pedestrians to walk safely even enter to crime hotspot location. Examining the path finding which is a typical function, technologies have been emerged beginning with the shorted path finding considering an initial distance, the fastest path finding considering speed limits and classes of roads, the free road path finding

considering a toll, and the optimum (safety) path finding with the real-time environment information.

Recently, demands are focused to guide roads for pedestrians are increasing due to the popularization of portable mobile devices such as smartphones, PDA, PMP (Figure 1).

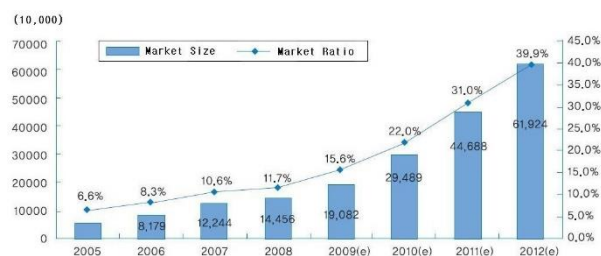


Fig.1 Trends and changes of smartphone markets (2009, Gartner)

All mobile devices nowadays can detect an IP address when someone surf an internet. This IP address can tell us someone location which clearly help us in this crime

The research using grant vot 59289 of Fundamental Research Grant Scheme (FRGS).

investigation. From these, we can calculate for the optimum path based on someone location. Therefore, it is required data and application programs specialized for pedestrian services. This clearly shown to us that crime prevention is not role from police only but rather seek help from others such as telecommunication company.

Based on characteristic of SCP itself, opportunity of a crime can be related to cost benefits, socioeconomic status, risk of detection, dependent on situational context, type of offence and access to external benefits. In addition, opportunities are dependent on the individual's current surroundings and consequential factors.

Emotions also play an important role in order to make this crime was failed or successful. It has three (3) important roles. First the people's state of emotionality is an important context on which rational conduct rests [4]. Second is the "sneaky thrill" of minor property crime also might operate more generally such that the anticipated emotional consequences of criminal conduct is one of the benefits or utilities ("thrills") that are weighed in the process of rational decision making [5]. Third as a sizable amount of research can attest, the anticipated emotional costs associated with criminal behavior might serve to effectively reduce the likelihood of such behavior [6]. Emotions are a central part of the psychological process of motivation as they heighten the saliency of certain desires, wants, and outcomes and thus energize people to pursue them [7]. If an victims gets panicked easily it is highly probable they won't think twice about making a wrong decision that will risk their life than an individual who is level headed.

This study analyzes factors having an influence on safety walking for data of environment setting constructing up to obstacles where pedestrians could pass through, and selects factors to be used for finding costs. Then, weights are assigned to the selected factors through the AHP (Analytic Hierarchy Process) method, and a path finding score is calculated using them. The path finding score is used to calculate a final path finding score for pedestrians through operations with a distance which is the existing finding cost.

Chapter 2 stated definitions for each keyword means. Chapter 3 derives factors having an influence on safety walking through an analysis on the existing studies. Chapter 4 proposed integrating AHP and Google Map, and Chapter 5 carries out detailed architecture of research framework. Finally, Chapter 6 discusses about conclusions and the future study subjects.

II . DEFINITIONS OF EACH KEY WORD

They have several keywords that readers will see in next section. To make understanding going smoothly well, this section will elaborate some keyword meaning in detailed.

A. Tactical Path Finding

In every company in this world have their unique business intelligence (BI) that drove to successful and profitable company. This business intelligence is made up from two (2) important components which are strategic plan and tactical plan.

Strategic plan or strategic intelligence is future-oriented, allowing a company to make educated decisions regarding future conditions in its particular marketplace or industry. It also permits the company's decision makers to visualize the future direction of the business and helps it recognize emerging trends and patterns within the particular industry and subsequently predict potential problems that may affect the current operating environment. In crime field, this strategy plan can included reducing the rewards that come from committing a crime, increasing the risk associated with offending, removing excuses for offending behavior and so on.

Whereas tactical plan means practical way that implemented to make sure our strategic plan is achieved. Tactical information was looked as involves a thorough and systematic analysis of current and emerging crime problems such as their causes and risk factors that is based on accurate, wide-ranging sources of information and has analysts with the capacity to interpret the data. In this research, tactical attributes are seemed important to bring our goal to tactical path-finding. Distance, visibility and obstacles can classify to tactical attributes.

B. Situational Crime Prevention

Situational Crime Prevention (SCP) is a prevention act that took opportunity as core elements in this area. It is chosen because this SCP has rarely been accorded attention in policy debates about crime control and also whether offenses are carefully planned or fueled by hate and rage, they are all heavily affected by opportunity [8]. This is strong by [9] that mentioned SCP capabilities are to change criminals' ideas about whether they can get away with a particular crime when opportunity are well manipulated by victims.

In real, SCP means to modify contextual factors to limit the opportunities for offenders to engage in criminal behavior [10]. It also involves identifying, manipulating and controlling the situational or environmental factors with certain types of crime [11]. This theory will be clearly understood when we go through to chess simulation study.

C. Analytic Hierarchy Process

The analytic hierarchy process (AHP) is a structured technique for organizing and analyzing complex decisions, based on mathematics and psychology. It was developed by Thomas L. Saaty in the 1970s and has been extensively studied and refined since then.

It has particular application in group decision making [12], and is used around the world in a wide variety of decision situations, in fields such as government, business, industry, healthcare, and education. Rather than prescribing a "correct"

decision, the AHP helps decision makers find one that best suits their goal and their understanding of the problem. It provides a comprehensive and rational framework for structuring a decision problem, for representing and quantifying its elements, for relating those elements to overall goals, and for evaluating alternative solutions.

Users of the AHP first decompose their decision problem into a hierarchy of more easily comprehended sub-problems, each of which can be analyzed independently. The elements of the hierarchy can relate to any aspect of the decision problem—tangible or intangible, carefully measured or roughly estimated, well or poorly understood—anything at all that applies to the decision at hand.

Once the hierarchy is built, the decision makers systematically evaluate its various elements by comparing them to one another two at a time, with respect to their impact on an element above them in the hierarchy. In making the comparisons, the decision makers can use concrete data about the elements, but they typically use their judgments about the elements' relative meaning and importance. It is the essence of the AHP that human judgments, and not just the underlying information, can be used in performing the evaluations [13].

The AHP converts these evaluations to numerical values that can be processed and compared over the entire range of the problem. A numerical weight or priority is derived for each element of the hierarchy, allowing diverse and often incommensurable elements to be compared to one another in a rational and consistent way. This capability distinguishes the AHP from other decision making techniques.

In the final step of the process, numerical priorities are calculated for each of the decision alternatives. These numbers represent the alternatives' relative ability to achieve the decision goal, so they allow a straightforward consideration of the various courses of action.

D. Chess Simulation Study

Chess game is widely well-known game and its characteristic more towards many-to-many attributes meanwhile in our focus research is more towards to one-to-many approach. In chess environment, each pieces will play a role in strategy games whether defense or attack enemy pieces. But in crime environment, victims play an important role to manipulate the situation effectively.

When the chess pieces in dangerous state, we need to safe our pieces from died or can develop strategy to eliminate enemy. But problem comes, mostly human will face disruption feelings caused by increasing an adrenaline hormone and leads them to make a wrong decision. This wrong decision is a big factor why our games always lost. This issue is not done yet; chess features also have do not provide a guide tool for a novice player. So, the percentage to winning a game was limit or none.

Next, we try to mapping out between chess case study and real crime case study, and both of them are very close relationship each other. For example, novice player in chess, we can assume it as foreigner people who do not familiar with

path. It also happen to expert people about path but cannot choose path well when the original path are blocked by construction or other else. Without a proper tool, it was impossible for them to choose a best secure path when facing crime situation or enter crime hotspots.

From perspective of adrenaline hormone also, researcher already stated our psychology balance will disrupted around 50% when facing a real crime situation [14]. This will lead them to a wrong and risky decision that leads them to crime's victim. Media also always show a seriousness crime in television or newspaper which makes public peoples always feeling fear and haunted with their own imagination.

Finally, after we understand enough how human decision works, we will go through to choosing a best secure path. This choosing approach will touched about closed list and open list that always being used in path-finding approach. In normal situation, people always have eight (8) decisions to choose. This choosing path will determine by closed list and open list of tactical path-finding. Closed list are used as a slightly best secure path which means, this path may contribute more to secure path meanwhile open list is list that not be checked from algorithms yet. In this process to choosing a possible closed list are controlled by pruning method. This pruning method will makes our system works efficiently and ease.

III. EXTRACT FACTORS AFFECTING SAFETY WALKING

The safety walking factors are divided into physical environmental elements and human subjective elements. The physical environmental elements are the ones for walking facilities, which are used as indexes for evaluating the walking environment and the human subjective elements are ones that could be differently represented depending on individual preferences and walking purposes etc. Handy S. [15] proved that the physical environment was an influence factor of safety walking by conducting a survey targeting pedestrians in different criteria's after classifying the safety walking purposes into leisure and movement means, and Moudon A. V. [16] analyzed on an influence factor of walking and cycling with evaluation indexes used in each field. And Lee C. [17] derived the importance of elements evaluating the walking affinity through a statistical analysis. Seo. H. L. [18] classified elements having an influence on walking into street environment network environment, and regional environment to analyze them, and Park S. H. [19] carried out a study to make them as indexes. Hieronymus C. Borst [20] classified elements having an influence on walking of the elderly and infirm into 25 items to make them as indexes and used them in path finding, Lee J. E. [21] classified the walking influence elements into physical environmental elements, changes of direction, visual field and accessibility to use for path finding through arbitrary weights.

Through these existing studies and distributed questionnaires, we finally came out with the classification for safety path-finding criteria are distance between victim and

offender, visual field of offender and numbers of obstacles exist as physical environmental and finally derived sub-criteria as Table 1 whereas in Table 2 shows the abbreviation meaning that listed in Table 1.

TABLE 1 : Factors having influences on safety walking

Components	Description
Distance	C1,C2,C3,C4,C5,C6,C7,C8,C9,C10
Visibility	V1,V2,V3,V4,V5,V6,V7,V8,V9,V10
Obstacles	OS,OM,OH (Size) OL,OA,OML (Light Emission) OFM,OSM, ONM (Move) O1,O2,O3,O4,O5,O6,O7,O8,O9,O10

TABLE 2: Abbreviations meaning that listed in Table 1

Abbreviation	Meaning
C1-C10	Cell-1 till Cell-10
V1-V10	Visibility-10% till Visibility-100%
OS, OM, OH (Size)	Obstacles Small, Obstacles Medium, Obstacles Huge
OL, OA, OML (Light Emission)	Obstacles Less Light, Obstacles Average Light, Obstacles More Light
OFM, OSM, ONM (Ability to move)	Obstacles Fast Move, Obstacles Slow Move, Obstacles No Move
O1-O10	Obstacles-10% till Obstacles-100%

From the Table 1, we derived sub-criteria with quantitative measurement where crime hotspots are identified from 10 x 10 arrays. If pedestrians approach area where crime hotspot radius is 10-cell array from initial place, then system will alert pedestrian to make decision. Similar to visibility and obstacles criteria representatively. They will measure through quantitative scale. However for obstacles criteria, it have slightly different measurement since this criteria can be sliced to favorable characteristic such as size, light restriction, ability to move and number of obstacles. This four (4) characteristic will framed the final result for obstacles criteria. If score is high, means, these obstacle criteria was important though which give advantages to victim in order determine safety path-finding.

A. Distance

Distance is an excellent measurement to determine the successful crime happens. It also close related with our velocity value. If the direction or path have a constraint in the middle, the velocity maybe slow compare than no obstacles in

direction. The velocity also influenced by personal behavior such as heart rate, fitness, and strategy. This will guarantee that our path is varying from others. As logic, factor that lead to successful crime is the distance between offender and pedestrians was very near and numbers of obstacle is many so that criminal can hide himself from obstacle. Thus, in our research study, distance is assumed by number of cell between pedestrians and offender. If cell is below than 5, then the possibility to have a crime are higher otherwise if cell is above than 5, then the possibility to have a crime is lower because pedestrian can took alternative path to run from it compare cell below than 5. However, it still depends to obstacles and visibility of offender.

B. Visibility

One factor that led to lack of awareness is the visibility of human itself in tracking the existence of offender from behind. If criminal visibility already locked the position of pedestrians, then it will be advantages to criminal itself. However, visibility of human are defined as circular arc (geometry) where it just can view around 120 degree only. This limitation degree was give advantage to pedestrians to escape from visibility area of offender. Furthermore this visibility value also will disturbed by obstacles that constraint them.

C. Obstacles

The obstacle is a third factor for successful crime. Obstacles have two (2) main functions which are first, it can be a weapon for offender to hide in order ambush a pedestrians and the second one, obstacle can be advantages to pedestrians to limit the visibility of offender and be a protection from criminal's attack. However the characteristic of obstacles is difference since we can categorize it into four (4) characters which are size, light emission, ability to move and numbers of obstacle. The description for this character can view in Table 1.

IV. THE PROPOSED INTEGRATING AHP AND GOOGLE MAP

Based on the observation made in previous section, we present the idea of tactical path-finding and IP address where we can use Google API to achieve this goal. The proposed approach framework is designed in a simple figure to make it clear.

Figure 2 shows the structure of proposed approach which consists of a data collection, data analysis from the distribution questionnaire, applying MCDM/AHP technique and integrating AHP and Google Map. Each phase contributed the output with each output from previous phase will then lead to the next phase. Finally, integration between criteria weights and maps is accomplished producing the suitability maps which have the potential area for crime.

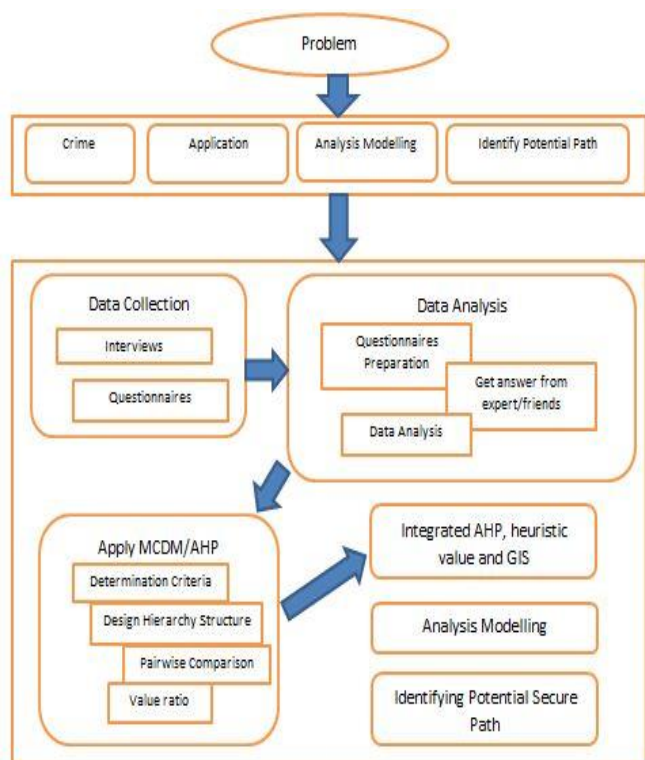


Fig. 2 Proposed Approach

A. Study Area

This study was conducted in Gong Badak, Kuala Terengganu. It is regional municipality located on the east side of Peninsular Malaysia and has population around 50,000 persons. Population is expected to reach 100,000 persons by 2020. Figure 3 shows the location of study area. Population increased occurred without proper controlling and monitoring has lead to several of problems such as criminal activity [22].

This region is the university area and busy street intersected by other university which is Universiti Sultan Zainal Abidin (UnisZA), where is the place with large pedestrian volume of people using surrounding facilities and willing to use public transportation.

From this, we can saw many crime hotspots that are spotted from PDRM file from January 2014 until December 2014. Most popular crimes that's happen was snatching criminal. It is because not 100% students are from Terengganu. Almost 70% are from outside Terengganu which is not very familiar with this place. So, with aid from our system, it could help students in taking a right path so that they can avoid from being crime victims.

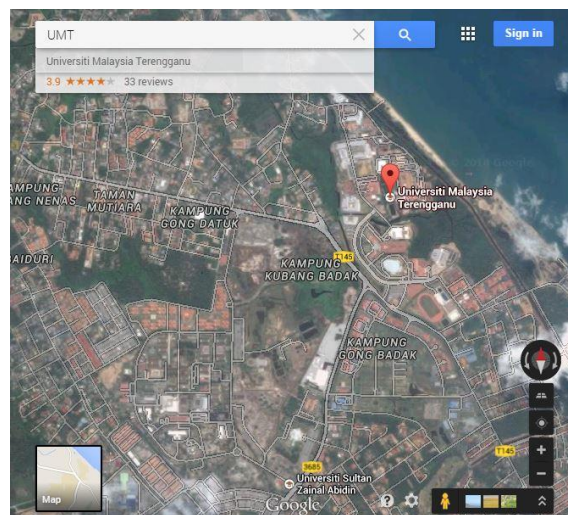


Fig.3 Universiti Malaysia Terengganu (UMT)

B. Data Collection

Spatial data were obtained from PDRM supported by Universiti Malaysia Terengganu (UMT). Then, layers of the selected area were generated and new layers were created using Google Map as the software to do the operations.

V. ARCHITECTURE OF PROPOSED FRAMEWORK

This section briefly describes the detailed architecture in three phases which are crime hotspots radius phase, criteria evaluation phase, and mapping phases.

C. Crime Hotspots Radius Phases

This is first phase in order to alert citizen about near crime hotspots between them. In here, we will use collision detection technique which modelled victim and crime hotspots respectively. This technique is widely used in games application. It is important to know when objects in a game collide.

There are a lot of ways that collision detection can be done. The most obvious method is to take every single vertex in a model and check to see if any of them are inside of another model. This method, however, is very time consuming to do, especially if have lots of models, or models with a lot of detail in them. To deal with this problem, game programmers will use an approximation of the model that is easier to check for collisions. The two methods that are most commonly used are bounding boxes and bounding spheres. With these methods, it basically builds a box or sphere around a model that completely covers the model. There will obviously be area that is outside of the model, but still inside of the bounding box or sphere, but an ideal bounding region will limit this as much as possible.

In this study, we would like to use circle rather than box since it give more precise information. With circle models, it can give information about center of circle and radius of circle very well. The basic concept is to figure out how far the point is away from the center of the circle. If that distance is less than

the circle's radius, the point is in the circle. If it's more than the radius, it's outside of the circle.

Finally, in short, we know that two circles must have some overlap if the two centers are closer together than the sum of their radii (radiuses). If the distance is more than the two radii combined, then there will be an empty gap between the two circles. If the distance is equal, the two circles are touching at one point (we'll call that an intersection). If the distance is less, then the two circles *must* be overlapping.

Finally, when both circles are overlapping, system will give alert to end-user for be careful from any possibilities until system give final result to choose secure path. The result of the process must pass through criteria evaluation first.

B. Criteria Evaluation

According to studies conducted from questionnaires, the most important criteria used to determine the potential area of crime in Gong Badak were recognized. Based on questionnaires, distance; visibility; and obstacles were chosen as the potential factors.

Every factor derived above has a qualitative property. In order to approach as the GIS methodology when utilizing them in path finding, however, a quantitative analysis is needed. This study uses the AHP method for such a quantitative analysis.

The AHP is a method that Thomas L. Satty devised in 1980, which is a multi-criteria decision making technique to select the optimum alternative by understanding the evaluation criteria and the alternative as a hierarchical structure. This AHP was developed based on the fact that the brain uses gradual or hierarchical analysis process when human beings make decision. According to results of the study, it is said that human beings follow three rules (setting of a hierarchical structure, setting of relative importance, maintenance of logical consistency) when they solve a problem.

After three (3) processes are done, it will come out with pathfinding scores. It will be varying in value because each attribute give different value. The indicator for best secure path is referring by high AHP value. In conclusion, two attributes likes obstacles and visibility give more advantageous if the value was larger compared than distance attributes where it give disadvantageous when value are higher.

C. Mapping Phases Using Google Map

In this study, layers overlay to raster conversion, clipping processes using Google Map API function and calculating criteria weight using an application based on AHP technique makes out the manipulation of this study. Using Google Map capabilities, criteria maps were converted to raster then they were classified into several classes. Finally, suitable map for potential path-finding crime area will be generated. This result will present a rank of best path to worse path. Then, suitability classification is divided into three classes to get the accurate result.

VI. DISCUSSION AND FUTURE WORK

This study extracted a variety of factors having an influence on safety walking in order to present a path finding method suitable to pedestrian path guidance which is recently issued. In addition, weights were calculated using the AHP to calculate the importance between the extracted factors, and these weight were used to assign scores of factors for each score. The logical validity was also proved using the consistency index when calculating weights for each factor through the AHP method. Finally, the finding cost was calculated for the pedestrian path finding by calculating these scores with distance values of the road network. The calculated finding cost applied to the road network in university busy street with large pedestrian volume, as a result, it was represented the result different from the path finding through a simple distance. It is analyzed because more suitable safety walking path was presented considering a variety of factors having an influence on safety walking.

However, a quantified verification for this is needed because it is only subjective view. In addition, a detailed study is needed for the criteria to distinguish attributes and the assigned scores also in the score assigning process.

The main cause affecting pedestrians in safety walking has many subjective elements. In other words, some people want a path having poor walking environment but faster way; on the other hand, some people want a walking path with longer way around but comfortable environment. The AHP method presented in this paper could be utilized for such personalized services. The AHP for decision making of public purposes should gather the major opinion, however for personal purpose such as safety walking, weights could be calculated according to a personal preference. Therefore, it could be utilized for personalized services through personal computing devices.

ACKNOWLEDGMENT

The authors would like to thank for continuous supports given by ACP Wan Abdul Aziz Wan Hamzah, Ketua Jabatan Siasatan Jenayah, Kuala Terengganu Royal Police Malaysia (RPM). This work was supported by grant from Fundamental Research Grant Scheme (FRGS) with vot number 59289.

REFERENCES

- [1] Program Transformasi Kerajaan, Laporan Tahun 2011, "Mengurangkan Jenayah Melalui Pelan NKRA", m/s 52-96.
- [2] Vienna. "Prevention: An Effective Tool to Reduce Corruption". Global Programme Against Corruption. Page 1-38.
- [3] Margaret Roper et al. "Khulisa Community-Based Crime Prevention Programme in Kwazulu-Natal 2006. Page 1-46.
- [4-6] Simpsons, S. (2000). Of crime and criminality: The use of theory in everyday life. Thousand Oaks, CA. Pine Forge Press.
- [7] Kaufman, B. (1998). Emotional arousal as a source of bounded rationality. *Journal of economic behavior and organization*, vol. 38, page 135-144.
- [8] V. Clarke, R., "Situational Crime Prevention – Successful

- Case Studies 2nd Edition. Library of Congress Cataloging, 1997: p. 1-43
- [9-11] Situational Crime Prevention Theory. URL: <http://crimeprevention.rutgers.edu/topics/SCP%20theory/theory.htm>
- [12] Saaty, Thomas L. Peniwati, Kirti (2008). Group Decision Making: Drawing out and Reconciling Differences. Pittsburgh, Pennsylvania: RWS Publications. ISBN 978-1-888603-08-8
- [13] Saaty, Thomas L. (June 2008). "Relative Measurement and its Generalization in Decision Making: Why Pairwise Comparisons are Central in Mathematics for the Measurement of Intangible Factors – The Analytic Hierarchy/Network Process". 102(2): 251-318. Doi:10.1007/bf03191825.
- [14] Cindy Dietrich. (2010). "Decision Making: Factors that Influence Decision Making, Heuristic Used, and Decision Outcomes. Vol.2 No. 2 page 1-3. URL: <http://www.studentpulse.com/articles/180/decision-making-factors-that-influence-decision-making-heuristics-used-and-decision-outcomes>
- [15] Handy, S. 1996 "Methodology for Exploring the link between urban Form and Travel Behavior." Transportation Research Part D 1(2): 151-165
- [16] Moudon, A.V. et al., 2003 "Walking and Bicycling: An Evaluation of Environmental Audit Instruments" American Journal of Health Promotion v.18 n.1 21-37
- [17] Lee, C. et al. (2006) "The 3Ds + R: Quantifying land use and urban form correlates of walking" Transportation Research Part D11 204-215
- [18] Hyeyoung Kim et al 2006, "An Accessibility-incorporated Pedestrian Routing Algorithm", Fall Conference on Geographic Information System Association of Korea 87-96
- [19] So-hynn Park et al, 2008, « Measuring Walkability in Urban Residential Neighborhoods:
- [20] Hieronymus C. Borst et al 2009, "Influence of environmental street characteristics on walking route choice of elderly people", Journal of Environmental Psychology 29, 477-484
- [21] Jongeon Lee et al 2008, "Department of Transportation Algorithm for Pedestrian in Shopping Area", Journal of Korean Society of Civil Engineers v.28 n.2D 147-154
- [22] Keith Harries. " Property Crime and Violence in United States: An Analysis of the Influence of Population Density". International Journal of Criminal Justice Sciences. Vol.1 Issue 2 July 2006.

Extended Cavity Model to Analysis Tunable Circular Disk Microstrip Antenna Using Genetic Algorithm

Sami BEDRA¹, Tarek FORTAKI², Siham BENKOUDA³, Abderraouf MESSAI³
¹Industrial Engineering Department, University of Khenchela, 40004 Khenchela, Algeria
²Electronics Department, University of Batna, 05000 Batna, Algeria
³Electronics Department, University of Constantine1, 25000 Constantine, Algeria
bedra_sami@yahoo.fr

Abstract—In this paper, the cavity model for simple circular disc microstrip antenna is extended with some modifications for the tunable geometry taking into account the anisotropy in the layer. The numerical results show that there are substantial deviations in calculated resonant frequency when substrate dielectric anisotropy is considered. Furthermore, significant variations are seen in the radiation patterns of the structures due to substrate anisotropy. Finally the effect of inclusion of air gap layer inserted between substrate and ground plane on the resonant characteristics is also investigated for fundamental and higher order modes.

Keywords—cavity model; genetic algorithm; anisotropic substrate; adjustable air gap.

I. Introduction

Microstrip antennas are becoming increasingly popular since they have small volume and a low-profile planar configuration. Easy mass production of such antennas using printed circuit technology leads to low fabrication cost. They are much easier to be integrated into microwave circuits on the same substrate. Especially, they can be made conformal to the host surface [1].

Some dielectric substances exhibit anisotropy due to their natural crystal structures or as the result of their production processes. Isotropic substances may also exhibit anisotropy at high frequencies. In the design of microwave integrated circuit components and microstrip antennas, anisotropic substances have been increasingly popular [2-10]. Uniaxial substrates have drawn more attention due to their availability such as sapphire, boron nitride and E-10 ceramic-impregnated Teflon. Their main drawback is narrow bandwidth characteristics, which is considerably avoided by operating the antenna around the resonant frequency. As an alternative, double-layered structure with air gap having adjustable thickness between the substrate and the ground plane is also found to be useful in obtaining the wide band operation. For both single and double-layered structures, accurate computation of resonant frequency is an important task and takes considerable interest in literature by various authors depending on the usage of various methods and approximations [11–19]. In this study, resonant frequency of double layered circular patch microstrip antenna is accurately determined via cavity analysis, using a simple effective

permittivity and patch radius expressions including modal effects. The aim of this work to perform an accurate and efficient analysis of circular-disc microstrip antennas on double layer, as well as to perform the analyses for circular microstrip antennas on a single layer substrate and on a tunable substrate, as particular cases.

II. Antenna Configuration and Design

The tunable circular microstrip antenna structure is shown in Fig.1. The resonant behavior of the antenna is independent of the feed so that the feed was not taken into account in the analysis.

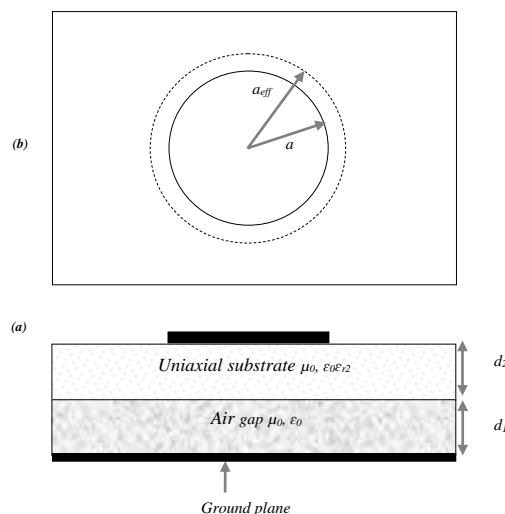


Fig. 1. Geometry of a circular microstrip antenna with air gap.

- a- Side view
- b- Top view

The Resonant frequency of this antenna can be determined from cavity model for various operational modes and structural parameters using proper equivalent model with effective structural parameters [3]. For this purpose, various expressions for effective patch radius a_{eff} and effective relative permittivity ϵ_{eff} are defined in literature [13, 15, 18, 19].

In this study, effective patch radius expression to approximate the modal effects is taken for the double-layered antenna in the modified form:

$$f_r = \frac{\chi_{nm} v_0}{2\pi a \sqrt{\epsilon_{req}}} \quad (1)$$

where χ_{nm} is the m th zero of the derivative of the Bessel function of order n , the value of which ($\chi_{01}=3.832$, $\chi_{11}=1.841$, $\chi_{21}=3.054$, $\chi_{31}=4.201$) determines the lowest and higher order modes as TM_{11} , TM_{21} , TM_{01} , and TM_{31} modes.

v_0 is the velocity of light in free space, a is the patch radius, and ϵ_{req} is the substrate relative permittivity of the equivalent structure which can be determined from the cavity model [14]

$$\epsilon_{req} = \epsilon_{r2}(d_1 + d_2)/(\epsilon_{r2}d_1 + d_2) \quad (2)$$

To account for the fact that small fraction of the field exists outside the dielectric; it is customary to use effective permittivity ϵ_{eff} in place of ϵ_{req}

$$\epsilon_{eff} = \epsilon_{req} - 0.9\epsilon_{req} \left[\frac{2d}{a} + \left(\frac{d}{a} \right)^2 \right] \quad (3)$$

Where $d=d_1+d_2$ and, ϵ_{r2} is the relative permittivity of dielectric substrate.

If we want to take the substrate uniaxial anisotropy's into account, the relative dielectric permittivity ϵ_r will be replaced with the tensor $\epsilon_r = \text{diag}(\epsilon_x, \epsilon_x, \epsilon_z)$ where ϵ_x and ϵ_z are the relative dielectric permittivity along x and z axis, respectively

- For the case of isotropic substrate with air gap, we use the effective dielectric constant ϵ_{eq} given in Eq. (2).
- For the case of uniaxially anisotropic substrate without air gap, ϵ_{eq} given in [21] Eq. (2) is used to determine d_e , there resulting values are:

$$\epsilon_{req} = \epsilon_z \quad (4)$$

$$d_e = d \sqrt{\frac{\epsilon_x}{\epsilon_z}} \quad (5)$$

To account fringe field effects, the circular patch radius a given in Eq. (1) should be replaced by its effective value [20, 21]. In this letter, a new effective patch radius expression is presented to compute the resonant frequency of a circular MSA with thin and without air gap for providing better accuracy. By utilizing the experimental data reported elsewhere [22-27], after many trials, the following model, depending on ϵ_{eff} , a and d , which produces good results, was chosen

$$a_{eff} = a + \left[\beta_1 + \left(\frac{\beta_2}{\epsilon_{eff}^{\beta_3}} \right) \right] d + \left(\frac{\beta_4}{a} \right) d^2 \quad (6)$$

where the unknown coefficients are determined by a genetic optimization algorithm. It is evident from (6) that the effective patch radius, a_{eff} is larger than the physical patch radius, a , provided the conditions and are satisfied. In the following section, the genetic optimization algorithm used in this work is described and then the application of the genetic algorithm to the problem is explained.

III. Genetic Algorithm

The GA [28, 29] is based on the evolution theory where weak species face extinction but strong ones survive and pass their genes to the next generation. However for the strong species to survive there is also a requirement for random injection of genes. As GA mainly manipulates matrices it is normally implemented using Matlab software. The step by step procedure of generating the software program is shown below.

Step 1: Each variable is assigned a number of binary digits so that the required accuracy of this variable is obtained in the final solution.

Step 2: All the variables in their binary form are grouped into a string which is called a chromosome.

Step 3: Matlab is used to select a fixed number of random chromosomes called a population out of all possible number of chromosomes that are present. This is called the current generation.

Step 4: Converting the digital value of each variable in a chromosome to an analogue value, the objective function (F) is evaluated and the relative fitness of each chromosome (P_i) determined. This relative fitness is defined as:

$$F = \sum_{i=1}^n \text{eval}_i [P_i] \quad (7)$$

Step 5: The selective probability is determined by:

$$P_{si} = \frac{eval_i[P_i]}{F} \tag{8}$$

The cumulative probability of the chromosomes is given as:

$$q_i = \sum_{j=1}^n P_{sj} \tag{9}$$

Then a random number 'r' is generated in the range 0 to 1. If $q_{i-1} \leq r \leq q_i$ then select P_{si} .

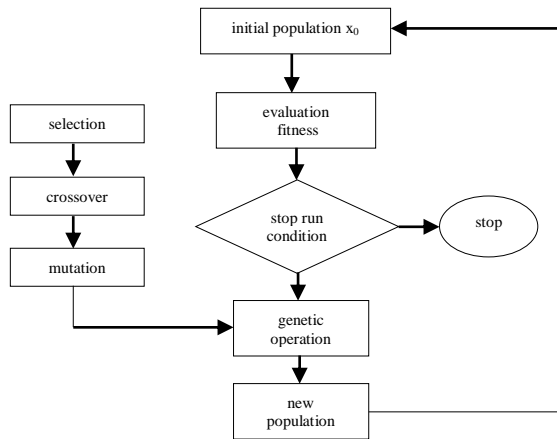


Fig. 2. Flow chart of genetic algorithm.

Step 6: Crossover is applied for random chromosomes between the parent and next generation to produce new off springs.

Step 7: The population is mutated by changing in a random way the value of the genes with the least significant bit having the highest probability of mutation and the most significant the least. The flowchart of GA is shown in Figure 2.

The next generation now becomes the parent generation and the above process is repeated until the genetic variation in the population is below a certain threshold.

As the number of generations increases both the cross over rate and the mutation rate are gradually reduced.

where $\beta_1, \beta_2, \beta_3$ and β_4 are given in the above equation are the coefficients to be determined by GA so as to minimize the following total absolute errors (TAE)

$$TAE = \sum |f_{me} - f_{ca}| \tag{10}$$

where f_{me} and f_{ca} are, respectively, the measured and calculated resonant frequency of circular MSA.

The control parameters in the optimization are as follows:

Maximum number of generations: 50

Population size: 100

Probability of 0 in initial population: 0.15

Probability of crossover: 0.8

Probability of mutation: 0.05

Optimization object: resonant frequency

TABLE I. RESULTS AND COMPARISON OF THE RESONANT FREQUENCIES OF MEASURED AND CALCULATED FOR THE FUNDAMENTAL MODE TM11 OF A CIRCULAR ANTENNA AND THE NO GAP CASE.

Physical and Electrical Parameters			Measured f_r (GHz)	Calculated Frequencies f_r (GHz)			Our results (GHz) f_r	Measured By
d (mm)	ϵ_{r2}	a (mm)		[23]	[26]	[27]		
1.588	2.5	34.93	1.57	1.592	1.555	1.559	1.557	[23]
3.175	2.5	34.93	1.51	1.592	1.522	1.529	1.522	
2.35	4.55	49.5	0.825	0.832	0.827	0.827	0.824	[26]
2.35	4.55	29.9	1.36	1.378	1.358	1.360	1.355	
2.35	4.55	20	2.003	2.060	2.009	2.012	2.007	
2.35	4.55	10.4	3.75	3.962	3.743	3.737	3.750	
2.35	4.55	7.7	4.945	5.352	4.938	4.924	4.947	
1.5875	2.65	11.5	4.425	4.695	4.414	4.438	4.416	[22]
1.5875	2.65	10.7	4.723	5.046	4.724	4.750	4.723	
1.5875	2.65	8.2	6.074	7.297	6.049	6.086	6.034	

TABLE II. COMPARISON OF THE RESONANT FREQUENCIES OF MEASURED AND CALCULATED OF A CIRCULAR ANTENNA HAVING AN AIR GAP; $a = 50\text{mm}, \epsilon_r = 2.32, d_2 = 1.59\text{mm}$.

Mode TM_{nm}	d_1 (mm)	Measured f_r (GHz)	Calculated Frequencies f_r (GHz)					Our results (GHz) f_r
		<i>Dahele</i> [12]	<i>Aboud</i> [13]	<i>Gurel</i> [16]	<i>Guha</i> [17]	<i>HFSS</i> [31]		
TM_{11}	0	1.128	1.159	1.129	1.130	1.162	1.134	
	0.5	1.286	1.298	1.281	1.274	1.334	1.283	
	1	1.350	1.368	1.359	1.344	1.435	1.349	
TM_{21}	0	1.879	1.927	1.876	1.881	1.934	1.881	
	0.5	2.136	2.167	2.128	2.119	2.203	2.130	
	1	2.256	2.280	2.258	2.235	2.353	2.239	
TM_{31}	0	2.596	2.665	2.584	2.594	2.356	2.588	
	0.5	2.951	2.994	2.930	2.921	2.645	2.927	
	1	3.106	3.150	3.109	3.080	2.829	3.080	

The unknown coefficient values of the model given by (6) are optimized by the genetic optimization algorithm just described. The optimum values found are

$$\beta_1 = 0.12, \beta_2 = 2.54, \beta_3 = 3.65, \beta_4 = 0.23 \quad (11)$$

The effective patch radius expression, a_{eff} , is obtained by substituting the coefficient values given by (11) into (6).

IV. Results and Discussion

In order to determine the most appropriate suggestion given in the literature, we compared our computed values of the resonant frequencies for the fundamental mode of circular microstrip antenna with the theoretical and experimental results reported by other scientists [22, 23, 26, and 27], all of which are given in Table I.

In order to check the accuracy of the model for two-layered case, the results are compared with an experimental and theoretical values presented in the previous work [12, 13, 16, 17, and 30] in Table II.

Figure.3 show the resonant frequency against the air gap thickness for several radius values of the circular-disc patch. It is seen that the operating frequency increases with the air layer thickness for a given value of patch size. So, antenna tuning is possible by introducing the air gap without changing the antenna parameters.

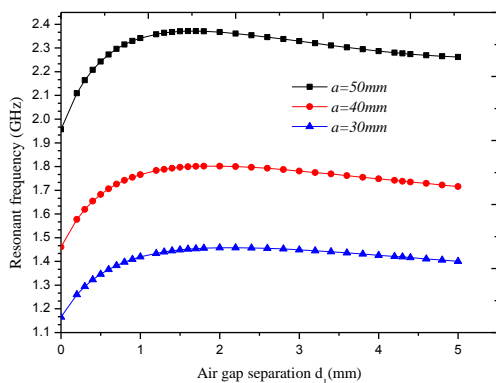


Fig .3. Resonant frequency versus air gap thickness for different values of patch radius, $d_2=1.27$ mm, and $\epsilon_x=\epsilon_z=2.32$.

It is observed that when the air separation grows, the resonant frequency increases rapidly until achieving a maximum operating frequency at a definite air separation d_{1fmax} . Note that the effect of the air gap is more pronounced for small values of d_1 show “Figure. 3”. When the air separation exceeds d_{1fmax} , increasing the air gap width will

decrease slowly the resonant frequency. These behaviors agree with those discovered theoretically for resonant frequency of circular patch antenna [16-19]. however, it depends inversely on the patch size for a given air gap width d_1 .

Next, the effect of uniaxial anisotropy on the resonant frequency is analyzed.

Fig. 4 depicts the influence of the patch radius on the resonant frequency of a circular microstrip antenna for anisotropic dielectric substrates (without air gap): Boron nitride ($\epsilon_x = 5.12, \epsilon_z = 3.4$). The substrate has thickness $d=1.27mm$. As it can be seen, the resonant frequencies reduce considerably with the dielectric substrates of Boron nitride.

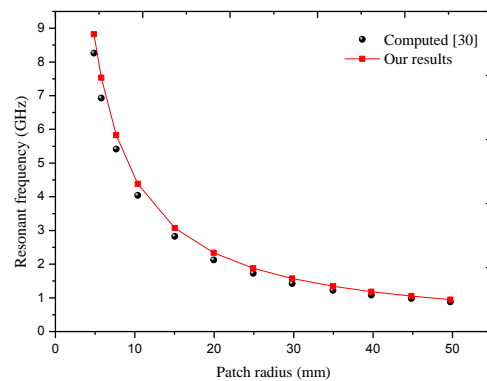


Fig. 4. Resonance frequency as a function of radius patch of a circular microstrip antenna without air gap on anisotropic substrate, ($\epsilon_x = 5.12, \epsilon_z = 3.4$), $d=1.27mm$.

Also it observed that the resonant frequency increases with the patch radius.

Fig. 5 depict the influence of the air gap thickness on the resonant frequency of a circular-disc microstrip patch for three anisotropic dielectric substrates: Boron nitride ($\epsilon_z = 3.4, \epsilon_x = 5.12$), Epsilam-10 ($\epsilon_z = 10.3, \epsilon_x = 13$), and Sapphire ($\epsilon_z = 11.6, \epsilon_x = 9.4$). The substrate has thickness $d_1 = 1.27$ mm and the air gap width is varied from 0 mm to 5 mm.

As it can be seen, the resonant frequency reduces considerably when the dielectric substrate changes from Boron nitride to Epsilam-10, and this is in contrast to what happens when the medium changes from Epsilam-10 to Sapphire. The obtained results show that when the permittivity ϵ_z is changed and ϵ_x remains constant, the resonant frequency changes drastically, on the other hand, we found a slight shift in the resonant frequency

when the permittivity ϵ_x is changed and ϵ_z remains constant. These behaviors agree very well with those reported by [6]. Also it is observed that the resonant frequency increases with the air gap thickness.

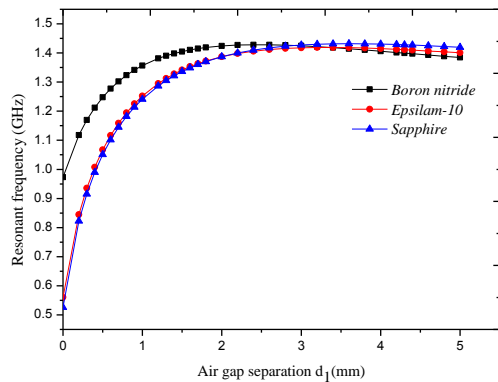


Fig. 5. Resonant frequency versus air gap thickness for different anisotropic dielectric substrates.

V. Conclusion

In this paper, a simple CAD formulation is presented based the cavity model for simple circular disc microstrip antenna is extended with some modifications for the tunable geometry taking into account the anisotropy in the layer. Computations show that the air separation can be adjusted to have the maximum operating frequency of the antenna. Extreme care should be taken when designing a microstrip antenna with thin air gap; since small uncertainty in adjusting the air separation can result in an important detuning of the frequency. The effects of a uniaxial substrate on the resonant frequency of structures are considered in detail. The results of the study will also be useful in the microstrip disk antenna design using uniaxial metamaterials.

References

- [1] Z.-S. Duan, S.-B. Qu, Y. Wu and J.-Q. Zhang, "Wide bandwidth and broad beamwidth microstrip patch antenna", *Electron. Lett.*, Vol. 45 No. 5, 2009.
- [2] N. G. Alexopoulos, "Integrated circuit structures on anisotropic substrates," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-33, pp. 847–881, 1985.
- [3] N. G. Alexopoulos and S. A. Maas, "Characteristics of microstrip directional couplers on anisotropic substrates," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-30, pp. 1267–1270, 1982.
- [4] E. Drake, R. R. Boix, M. Horno, and T. K. Sarkar, "effect of substrate dielectric anisotropy on the frequency behavior of microstrip circuits", *IEEE Trans. Microwave Theory Tech.*, vol. MTT-30, pp. 1267–1270, 1982.
- [5] H. Y. Yang and N. G. Alexopoulos, "Uniaxial and biaxial substrate effects on finline characteristics," *IEEE Trans. Microwave Theory Tech.*, vol. MTT- 35, pp. 24–29, 1987.
- [6] F. Bouttout, F. Benabdelaziz, A. Benghalia, D. Khedrouche and T. Fortaki "Uniaxially anisotropic substrate effects on resonance of rectangular microstrip patch antenna" *Electron. Lett.*, vol. 35, No 4, pp. 255-256, 1999.
- [7] A. Zhao, J. Juntunen, and A. V. Raisanen, "An Efficient FDTD Algorithm for the Analysis of Microstrip Patch Antennas Printed on a General Anisotropic Dielectric Substrate," *IEEE Trans. Microwave Theory Tech.*, vol. 47, NO. 7, 1999.
- [8] A. Luiz P. S. Campos and A. G. d'Assunçio "Hertz Vector Potential Analysis of FSS on Anisotropic Substrates" *Proceedings SBMO/IEEE MTT-S IMOC 2003.*
- [9] V. Losada, Boix R R, and Horno M. Full-wave analysis of circular microstrip resonators in multilayered media containing uniaxial anisotropic dielectrics, magnetized ferrites, and chiral materials," *IEEE Trans. Microwave Theory Techniques*, Vol. 48, 1057–1064, 2000.
- [10] C. S. Gurel, and Yazgan E. Characteristics of a circular patch microstrip antenna on uniaxially anisotropic substrate. *IEEE Trans. Antennas Propagat*, Vol. 52; 2532–2537, 2004.
- [11] K.F. Lee, K.Y. Ho, and J.S. Dahele, "Circular disk microstrip antenna with an air gap", *IEEE Trans Antennas Propag 32*, 880– 884, 1984.
- [12] J.S. Dahele and K.F. Lee, "Theory and experiment on microstrip antennas with air gap", *Proc IEE 132 (Part H)*, 455–460, 1985.
- [13] F. Abboud, J.P. Damiano, and A. Papiernik, "A new model for calculating the input impedance of coax-fed circular microstrip antennas with and without air gap", *IEEE Trans Antennas Propag 38*, 1882–1885, 1990.
- [14] J.S. Joy and B. Jecko, "A formula for the resonance frequencies of circular microstrip patch antenna satisfying CAD requirements", *Int J RF Microwave Comp Aided Eng.*, Vol. 3, 67–70, 1993.
- [15] K. Guney, "Resonant frequency of electrically-thick circular microstrip antenna", *Int J Electron.*, Vol. 77 , 377–385, 1994.
- [16] C. S. Gurel and E. Yazgan, "Resonant frequency of an air gap tuned circular disk microstrip antenna", *Int J Electron.*, Vol. 87, 973–979, 2000.
- [17] D. Guha, "Resonant frequency of circular microstrip antennas with and without air gaps", *IEEE Trans Antennas Propag.*, Vol. 49, 55– 59, 2001.
- [18] T. Gunel, "Continuous hybrid approach to the modified resonant frequency calculation for circular microstrip antennas with and without air gaps", *Microwave Opt Tech Lett.*, Vol. 40, 423– 427, 2004.
- [19] C. S. Gurel, E. Aydin, and E. Yazgan, "Modified resonant frequency calculation for two-layered circular patch microstrip antenna", *Microwave Opt Tech Lett.*, Vol. 49 , 2263–2267, 2007.
- [20] H.I. Kang and J.T. Song, "Electrically tunable rectangular microstrip antenna", *Electron. Lett.*, Vol. 146 No 7, 2010.
- [21] Y. Tighilt, F. Bouttout, and A. Khellaf, "Modeling and Design of Printed Antennas Using Neural Networks", *Int J RF and Microwave CAE.*, Vol. 21:228–233, 2011.
- [22] T. Itoh and R. Mittra, Analysis of a microstrip disk resonator, *AEU Int J Electron Commun.*, Vol. 27, 456–458, 1973.
- [23] J.Q. Howell, Microstrip antennas, *IEEE Trans Antennas Propagat AP-23*, 90–93, 1975.
- [24] S.A. Long, L.C. Shen, M.D. Walton, and M.R. Allerding, Impedance of a circular disk printed- circuit antenna, *Electron Lett.*, Vol. 14, 684– 686, 1978.

- [25] S. Yano and A. Ishimaru, A theoretical study of the input impedance of a circular microstrip disk antenna, *IEEE Trans Antennas Propagat AP-* 29, 77–83, 1981.
- [26] F. Abboud, J.P. Damiano, and A. Papiernik, “New determination of resonant frequency of circular disc microstrip antenna: application to thick substrate”, *Electron Lett.*, Vol. 24, 1104–1106, 1988.
- [27] Q. Liu and W.C. Chew, “Curve-fitting formulas for fast determination of accurate resonant frequency of circular microstrip patches”, *IEE Proc Microwave Antennas Propagat Pth* 135, 289 – 292, 1988.
- [28] Anirban Karmakar , Rowdra Ghatak , R.K. Mishra , D.R. Poddar, “Sierpinski carpet fractal-based planar array optimization based on differential evolution algorithm”, *Journal of Electromagnetic Waves and Applications.*, Vol. 29, 247-260, 2015.
- [29] Sharmin Shabnam, Suvrajit Manna, Udit Sharma, Pinaki Mukherjee., “Optimization of Ultra Wide-Band Printed Monopole Square Antenna Using Differential Evolution Algorithm”, *Advances in Intelligent Systems and Computing* Vol. 339, pp 81-89, 2015.
- [30] A. K. Verma, and Nasimuddin, “Analysis of circular microstrip patch antenna as an equivalent rectangular microstrip patch antenna on iso/anisotropic thick substrate,” *IEE proc. Microw. Antennas propag.* vol.150, No. 4, pp. 223-229, 2003.
- [31] HFSS: High Frequency Structure Simulator-Ansoft Corp.

Improving the Reuse of Services in Geospatial Applications with XMDD Technology

Samih Al-Areqi
Institute of Computer Science
University of Potsdam
Potsdam, Germany
samih@cs.uni-potsdam.de

Anna-Lena Lamprecht
Institute of Computer Science
University of Potsdam
Potsdam, Germany
lamprecht@cs.uni-potsdam.de

Tiziana Margaria
Institute of Computer Science
University of Potsdam
Potsdam, Germany
CSIS, University of Limerick, and
Lero, the Irish Software Research Center
Limerick, Ireland
margaria@cs.uni-potsdam.de

Abstract— In recent years, the geospatial application domain has embraced component-based development and service orientation to support software reuse. However, due to the specific characteristics of geospatial applications, caused by complex and comprehensive analysis processes and heterogeneous data, the reuse of services faces particular barriers in this domain. Providing application experts without a strong programming or technical background with simple means to reuse these services is an important challenge. This paper describes how we followed the eXtreme Model-Driven Development (XMDD) paradigm to improve the reuse of geospatial services, namely by (1) performing rigorous service abstraction of geospatial tools to be reused in large scale applications, (2) using the java electronic tools integration (jETI) technology for enabling the remote execution and integration of services, and (3) supporting service composition at the user level by using the java application building center (jABC) process modeling framework. Concretely, we discuss how we improved the reuse of services for the assessment of the impacts of sea-level rise.

Keywords— Service reuse, geospatial applications, scientific workflows, agile methodologies, extreme model-driven design.

I. INTRODUCTION

Building applications based on the reuse of existing components or services has noticeably increased in many domains. In the geospatial application domain, big geographic data, lack of interoperability, and complex analysis processes constitute barriers to ensuring a successful and wide reuse of components and services. Service-oriented architecture (SOA) principles and Web Service technology have been embraced by the geospatial domain and many works quickly followed the trend of building geospatial applications by reusing components and services. Several works focused on the construction of domain-specific applications by assembling

and reusing geospatial processes and data as services [1, 2, 3]. To facilitate the reuse of geospatial services, in the last decade many researchers followed the Open Geospatial Consortium's (OGC) Web Service standards [4] to build geospatial applications by composing services (eg. [5, 6, 7, 8, 9]). Workflow technologies such as jOpera have been applied early to the geospatial domain [10], and also the Kepler scientific workflow system [11] has soon been applied to handle distributed geospatial data processing using Web Services [12] and to compose OGC services [13,14]. Other works used BPEL-based business workflow technology to orchestrate geospatial services [15]. Nevertheless, learning how to apply these technologies to build a system based on

services remains complex for application experts, in particular with the interoperability challenges of geospatial data. A result from embracing service orientation in the geospatial domain is that the scientific data has become increasingly remotely accessible in a distributed fashion through standardized geospatial Web Services [2]. Thus, scientific communities become more aware of the benefits of sharing their data and computational services, and are thus contributing to distributed data and services. However, researchers should also not be too occupied with exploring how to reuse and compose geospatial services in order to develop own software applications tailored to their specific needs.

Despite substantial efforts by the OGC to provide standards for geospatial Web Services, turning spatial data and processes into loosely coupled components and interoperable geospatial services is suffering from the technical complexity of using the standards. In addition, there is a lack of a framework for facilitating service execution, thus users face a great challenge when it comes to *servification*, that is, the process of turning arbitrary software components into proper services. Attempts were made to improve the reuse of service in geospatial applications for end users (application experts), and some works addressed technical complexities of workflow systems by enabling Web-based workflow composition and editing [16, 12], while others proposed a model-driven way of geospatial Web Service composition [17]. Lately, cloud technology has been used to support efficient resource allocation and execution for scientific workflows [18, 19]. However, more technical efforts are required to handle the lightweight geospatial service execution in the cloud as described in [20].

The aim of this paper is to show how we follow the eXtreme Model-Driven Development (XMDD) paradigm [21] to achieve an improvement of geospatial services reuse. We do this by (1) performing *servification* of sea-level rise impacts analysis tools and data, (2) using the jETI technology for the remote integration and execution of the services, and (3) enabling users to compose services into workflows using the jABC framework. The rest of the paper is organized as follows: Section II gives an overview of component and service reuse, geospatial services, scientific workflow technology and agile method-ologies, in particular Section II-A introduces the jABC framework and Section II-B gives a summary about the jETI framework. Section III describes the proposed approach to address service reuse, which comprises *servification*, service execution, and service reuse. Finally, Section 0 discusses conclusions and plans for future work.

II. BACKGROUND

Software reuse ranges from simple functions to complete applications and is often considered the most effective means for improvement of productivity and maintainability in software development projects. The emergence of paradigms such as component-based software engineering (CBSE) and

service-oriented software engineering (SOSE) has leveraged the development of applications based on reuse of existing components and services. It significantly increased the possibilities of building systems and applications from reusable components [22]. CBSE aims at encouraging reuse of software applications, where systems are built by assembling components already developed and prepared for integration. In addition, it leverages the emergence of middleware technologies, such as object standards, to make software reuse a reality [23]. Although CBSE has proven to be successful for software reuse and maintainability, software developers are facing today more complexities, such as varying platforms, varying protocols, various devices, etc. [24].

Services are a natural further development of software components. They can be defined as loosely coupled reusable software components that encapsulate discrete functionality [25]. The paradigm of service-oriented software engineering overcomes the issues of heterogeneity and interoperability challenges of CBSE by defining standards to support easy service reuse and composition for system developers. Web Service standards are defined to represent computational or information resources that can be used by other applications. Service-oriented architectures (SOA) support distributed systems development based on service reuse. The major benefit from SOA standards (such as WSDL to describe services) is to enable interoperability across applications over different platforms.

The interoperability challenge in the geospatial domain and the advancements in general Web Service technologies and in GIS service standards such as Spatial Data Infrastructures (SDI) and OGC Web Service standards encouraged the migration from the traditional form of stand-alone geospatial applications to loosely coupled components, interoperable geospatial services, and grid computing. The OGC standards that are based on the service-oriented architecture have been designed to ease the reuse and integration of geospatial Web Services. However, they do not comply with the Web Service standards as defined by the W3C and OASIS. Therefore, developing geospatial services and composing them based on OGC standards requires additional technical efforts from both developers and users.

Scientific workflow technologies aim to facilitate and support the composition and execution of complex analysis processes in a flexible fashion [26]. In contrast to the communication- and document-oriented workflows in the business domain, scientific workflows are data- and computation-oriented. Despite their promise to simplify the service composition process, scientific workflow management systems are often inherently complex and challenging in use and design, especially where the managed resources are heterogeneous. Furthermore, many current workflow technologies are designed to support service composition at a lower, technical level, and not at a level where average users

can handle the composition and execution tasks. Composing services of geospatial applications in such workflows has a great focus on the data flow, and the underlying computation infrastructure has a major impact on the execution of the workflows. While clusters and grids are traditionally used to run large-scale scientific workflows, lately the trend is to execute the scientific workflows in hosting platforms such as clouds. Cloud computing "enables small and medium sized companies to deploy their Web-based applications in an instant scalable fashion without the need to invest in large computational infrastructures for storing large amounts of data and/or performing complex processes" [27]. Further, the use of VM images in the cloud to store computational environments and on-demand provisioning capabilities will improve reproducibility, which is significantly important for scientific workflows [28]. Users need however programming environments that support an easy design and execution of the scientific workflows.

A. jABC

Agile methods in the spirit of [29] have become increasingly popular in software development. Their core principle is to open software development to customers and users, in order to improve productivity, quality and stakeholder collaboration and satisfaction. The eXtreme Model-Driven Design (XMDD) paradigm [21] is an extremely rigorous way of model-driven development that supports a very agile and cooperative development of service-oriented systems by turning system development into user-centric orchestration of intuitive service functionality [30]. The multi-purpose process modeling and execution framework jABC [31] inherits the power of XMDD to enable end users to easily use and compose services into agile workflows. Its way of handling the collaborative design of complex software systems has proven to be effective and adequate for the cooperation of non-programmers and technical people. It enhances other modeling practices like the UML-based RUP (Rational Unified Process) and by leveraging plugin technology supports most activities needed along the development lifecycle like animation, rapid prototyping, formal verification, debugging, code generation, and evolution. In fact, compared with other workflow systems, the jABC offers a number of advantages that play a particular role when integrating off-the-shelf, possibly remote functionalities [32]:

- **Simplicity:** Focusing on application experts, who are typically non-programmers. The basic ideas of the modeling process have been explained in past projects to new participants in less than one hour.
- **Agility:** Models, and artifacts change over time based on expected requirements, therefore the process supports evolution as a normal process phase.
- **Customizability:** The building blocks which form the model can be freely renamed or restructured to fit the habits of the application experts.

- **Consistency:** The same modelling paradigm underlies the whole process, from the very first steps of prototyping up to the final execution, guaranteeing traceability and semantic consistency.
- **Verification:** With the model checking plugin, the jABC supports users to consistently modify their models. The basic idea is to define local or global properties that the model must satisfy and to provide automatic checking mechanisms.
- **Service orientation:** Existing or external features, applications, or services can be easily integrated into a model by wrapping the existing functionality into building blocks that can be used inside the models.
- **Executability:** The model can have different kinds of execution code. These can be as abstract as textual descriptions (for example in the first animations during requirement capture), and as concrete as the final runtime implementation.
- **Universality:** Based on Java as largely platform-independent, object-oriented implementation language, jABC can be easily adopted in a large variety of technical contexts and of application domains.

The service concept of jABC is very close to an intuitive understanding of service that is required to be ubiquitously accessible (location-agnostic) and mechanically configurable [33]. The term *service* is used to denote functional building blocks (SIBs), which are viewed as independent from their location, the program entity, and hardware-platform which provides them. The SIBs are orchestrated with their operational or behavioral semantics in mind. Concretely, this means that each SIB, once activated, executes its logic and upon termination triggers subsequent SIBs according to the outcome of this execution. This methodology of composition has been termed lightweight process coordination [9], focusing on operational aspects of the application rather than structural properties of the software. The notion of service in jABC is therefore fundamentally different from the Web Service notion. The ties to Web-communication protocols are not an essential part of jABC, but provided by the jETI technology [34]. The jABC process modeling and execution framework [31] has been applied to support agile workflows in different scientific applications domains in the last years, predominantly in the field of bioinformatics and for geospatial applications (cf. [35, 36, 37]). The framework has furthermore been extended by functionality for semantics-based semi-automatic service composition, which has been shown to be beneficial especially for dealing with variant-rich scientific workflows [35].

B. jETI

The Java-based jETI [34] is a redesigned version of the Electronic Tool Integration (ETI) [38] platform, an open

platform for the interactive experimentation with and the coordination of heterogeneous software tools via the internet. It was designed to provide:

- tool users with an instant hands-on experience with the tools, without need to download and install the software - which too often costs a considerable amount of effort and time, and
- tool providers with an environment where they may publish and promote their tools, making experimentation available to end-users without the burden and legal issues of direct distribution, and where they may receive valuable feedback.

Although the ETI platform offered a good solution to integrate software tools remotely, its servers were too complicated for both the tool providers and users. To follow the rapid development methodologies, the jETI framework overcomes these problems by applying newer technologies and standards that internally base on Web services and Java technology. It replaces the requirement of physical tool integration of the original ETI approach by very simple registration and publishing platform. Corresponding to the Web services functionality and service description standards such as WSDL, jETI uses an HTML tool configurator to create service descriptions. This allows providers to register a new tool functionality just by uploading the tool to the server and filling the description information (interface definition, input and output parameters, etc.) into a simple template form. All this information is internally maintained in an XML file and available for further use. For example, SIBs for use in the jABC framework can be generated automatically from the specifications, so that the services can easily be used within the jABC. Thus, with the lightweight remote service technology of jETI, users are able to

1. considerably simplify the integration process, and at the same time
2. flexibilize the distribution, version management and use of integrated tools,
3. broaden the scope of potential user profiles and roles from different application domains to solve complex problems and
4. solve the scalability problem connected with tool maintenance and evolution.

III. MAIN APPROACH

In this section we discuss how we used the jABC and jETI technologies to improve the reuse of geospatial services. As shown in Figure 1, the methodology involves three phases: First, the scientific tools (in this case tools for sea-level rise impacts analysis) which are used for geospatial applications, are servified (turned into services). Second, these services are reused to construct geospatial applications in the form of workflows, and finally the workflows (WF) are executed,

accessing the remote services. A concrete description of each phase is given in the following sections.

A. Servification

Several tools and applications have been developed to analyze the risk index of climate impacts, such as data creation, conversion, and visualization tools. The scientific tools that we used for our application address the analysis of the impacts of sea-level rise. These tools are used in the ci:grasp¹ climate information platform. They are based on scripts in the GNU R language that comprises several tools for spatial analysis. The srtmtools-package [39] used for the data analysis provides the methods required to produce results as presented on ci:grasp. It combines various tools that are based on different packages. For instance, a raster package tool² for data reading, writing, manipulating, analyzing and modeling of gridded spatial data, the Gdal tool³ for data conversion, and other packages for data visualization such as Png⁴ and plotGoogleMaps [40].

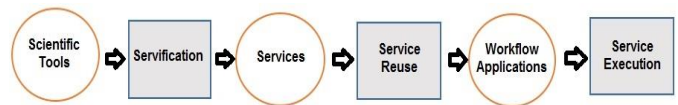


Fig 1. From geospatial tools to running workflows.

According to the service orientation paradigm, which postulates that any kind of computational resource should be seen and handled as a service – that is, a well-defined unit of functionality with a well-defined interface – to provide a high level of abstraction and reusability (cf., e.g., [41]), we use the term servification to refer to the process of turning arbitrary software components into proper services that are adequate, for example, for (re-) use in workflow management systems. Concretely, in the servification phase, the analysis processes of sea-level rise impacts implemented for ci:grasp and coded in R scripts have been decomposed into loosely coupled services. The decomposition handled service reuse by determining the most frequently used process steps in various applications of climate impact assessment and perform rigorous abstraction to ensure a great level of reuse for the services. Through jETI, a description for each service, equipped with well-defined inputs and outputs, is configured on the server and connected with the corresponding script file. After that, services are generated automatically into SIBs, so that they can easily be consumed by the jABC.

¹ <http://www.cigrasp.org>

² <http://cran.r-project.org/web/packages/raster/>

³ <http://www.gdal.org>, <https://r-forge.r-project.org/projects/rgdal/>

⁴ <http://www.rforge.net/png>

So far, 17 services for different data creation, computation, and data output tasks have been created (see Table II). Concretely, its three subclasses of SLR services concern: data creation (comprising 6 services), computation (6), and output (5). With regard to working with the jABC, this is the domain modeling phase, which enables us to model the domain of the sea-level rise example by integrating such created services and organize them in domain-specific taxonomies, so that they are ready for use in the actual workflow design phase. Figure 4 shows how the SLR services can be taxonomically classified and categorized into three groups:

- Data creation (loading, clipping, masking and converting data)
- Computation (of flooded areas, yield loss, caloric energy loss and land loss classes)
- Output generation (creation of PNG, PDF, TXT, GeoTiff/ASCII output files and result visualization in an interactive map)

by means of service compositions. This section demonstrates how the agile methodologies supported by the jABC framework make an essential contribution to increasing geospatial service reuse. Concretely, we will show how based on the newly created domain-specific services and the large library of SIBs for common functionality that comes with the jABC framework, we easily construct different workflows for SLR impact assessment in an agile workflow-based way.

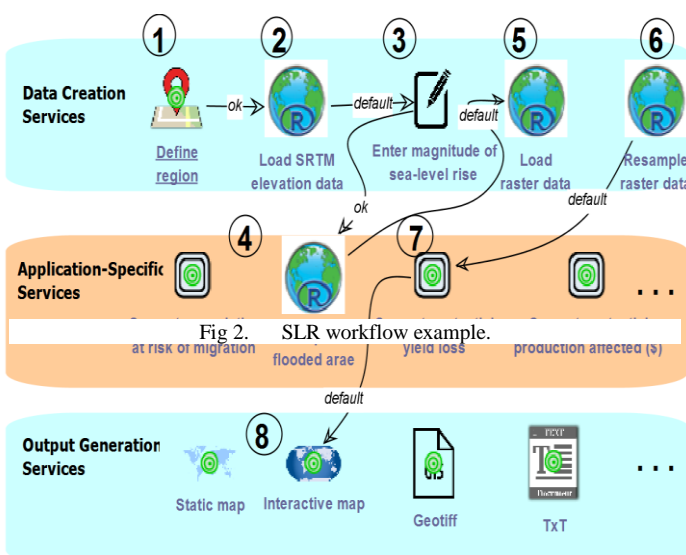
Figure 2 shows a simple workflow for assessing the impact of sea-level rise on the agricultural yield loss for a region to be selected by the user. From top to bottom, the services belong to three different groups of functionalities (data creation, application-specific computation services, and output generation). Starting in the upper left corner (the SIB with the underlined name denotes the starting point), the workflow performs (1) definition of the investigated area by coordinates of name; (2) downloading the digital elevation model of the selected area; (3) entering the magnitude of sea level rise; (4) computation of the flooded area; (5) load raster data from yield dataset; (6) resample two different data sets (in this example land loss data with yield data); (7) computation of the yield loss cause by the flooding; and (8) generation of an output file with results in an interactive Google map.

Table I. APPLICATION OBJECTIVES TO ASSESS SLR IMPACTS

Application	Description
compute rural and urban GDP at risk	focuses on potential economic damage in coastal communities
compute population at risk of migration	focuses on the number of people that would be affected
compute potential yield loss	compute potential production value affected in USD
compute potential land loss (ha)	determine the area that will be potentially inundated
compute potential production affected (\$)	focuses on the economic value of the agricultural loss
compute potential caloric energy loss	focuses on the potential number of peoples annual diets lost

B. Service Reuse

To increase service reuse, a significant aspect is to facilitate service consumption and to make the composition easy and flexible for a wide range of communities and people, so that the scientific community (e.g. geospatial application experts) can use and understand the service principles and build applications



Service S	No. Reuses of S	No. Workflows Using S
Data creation services		
Load raster data	28	11
Load SRTM data	11	11
Resampling	23	10
Clipping	6	6
Masking	8	8
ConvertKgTokcal	4	4
Computation services		
Compute flooded area	10	10
Compute land loss classes	1	1
Compute population at risk	2	2
Compute yield loss	8	8
Identify agriculture area	8	8
Identify flooded agriculture area	8	8
Output services		
Produce Pdf file	55	11
Produce image file	55	11
Produce Geotiff file	55	11
Produce text file	55	11
Generate interactive map	55	11

Table II. FREQUENCY OF SERVICE REUSE

In order to support the reuse of workflows, multiple abstraction levels have been introduced by making use of the hierarchical modeling capabilities of the jABC. Some of the

SIBs in the figure are marked by a green circle, which indicates that the functionality represented by this building block is actually more complex and defined by a separate (sub-) model. For example, SIB (7) encapsulates a (sub-) model for the computation of the yield loss (shown in Figure 3). Note that it again makes use of other (sub-) models, as the SIB to select potential yield data is a composite service that allows for the computation of several types of yield loss for different climate scenarios. This hierarchical modeling style allows to organize workflow applications at different levels of abstraction, from coarse-granular and more conceptual views at the higher levels, down to fine-granular and more technical views at the lower levels. The current SLR workflow scenario comprises six different computations (applications), as summarized in Table I.

According to different objectives to assess SLR impacts, each workflow application has several variations of workflow instances. For evaluating the reuse of geospatial services in the workflow variations, we used the jABCstats framework [42] to calculate the frequently of services reuse. Table II shows that the 17 created services as described and classified in section III-A have been reused 392 times in total, and within 11 workflow variations for sea level rise impact analysis. This also reflects that the services have contributed to a significant number of reuses in the different workflow applications. Not surprisingly, that data creation and output generation services are reused for all SLR applications. Figure 4 depicts the taxonomic classification and reuse levels of all services. Note that these services could also be reused in other analyses of climate change drivers included on ci:grasp, such as changes in temperature and precipitation and creased drought risk, and to other risk analyses related to climate impacts. Furthermore the services of data creation, resampling and output generation and visualization are more likely to be reused in the geospatial application domain in general.

C. Service execution

We believe that performing rigorous servification and providing an easy and flexible way to consume services in geospatial applications significantly improves their reuse. However, geospatial services deal with large data sets and need comprehensive computing resources.

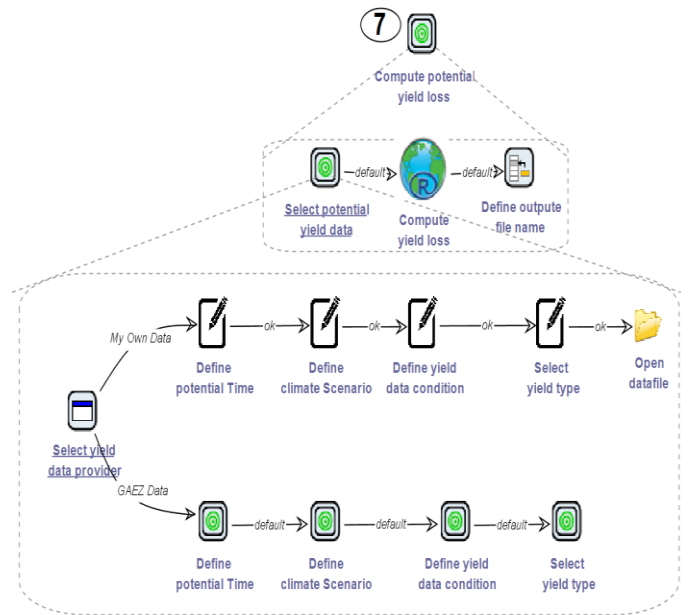


Fig 3. Computation of the yield loss.

In this section we show how jETI handles the remote execution of geospatial services. As mentioned in section III-A, the created services are based on several packages and use a diversity of data sets (e.g., elevation, land-use, population density or yield data). Consequently, these packages and data and the pre-configuration corresponding to the operating system platform are required to perform the execution of services. The jETI platform offers a lightweight remote component (tool) to further simplify integration and execution of software tools, it can be seen as a tool that enhances other tools and frameworks by the integration, organization and execution of remote functionalities, so that users do not have to deal with the required configuration to execute the services.

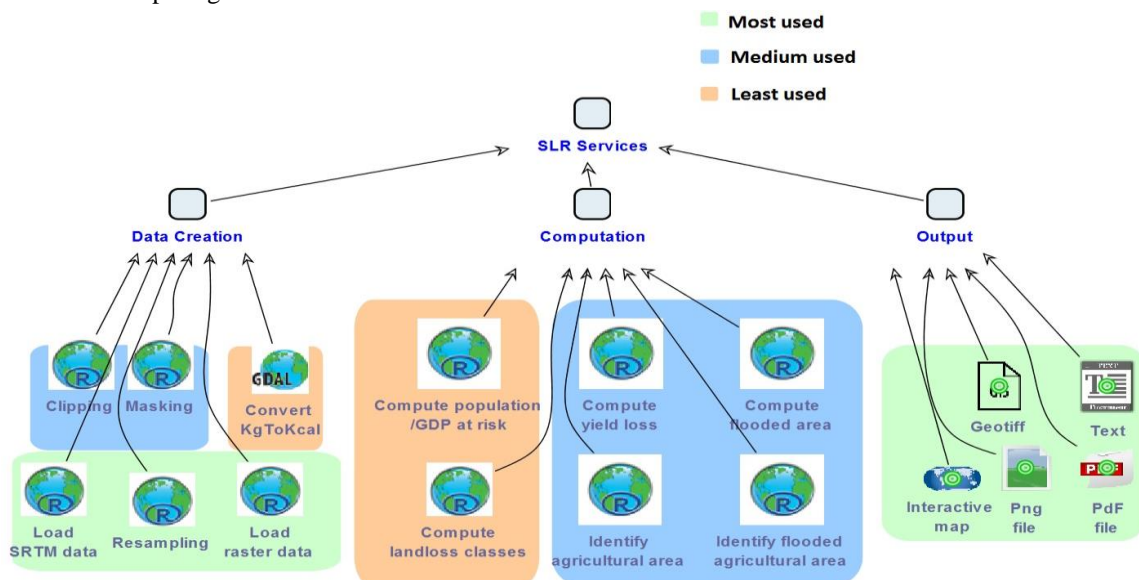


Fig 4. Service taxonomy.

In our case, we use the jETI server to support a convenient and flexible platform that enables users to execute geospatial services without dealing with the related configurations. On the jETI server, script files for created services are installed and wrapped to enable convenient automated invocation. The required configuration includes the installation of the GNU R language and packages such as Raster, Rgdal, ClassInt, Png and plotGoogleMapall. The jETI server itself runs in a virtual machine image based on a Debian Linux operating system. Managing the underlying infrastructure can be an issue as well, thus we follow the recent trend of using cloud technology to host our server and services. Thus, in our solution, the users design their workflow applications with the jABC, which during workflow execution submits jobs to the cloud where the jETI services for risk analysis of SLR impacts are hosted, as shown in Figure 5. This allows us to benefit from the advantages of cloud, such as resource scalability and data availability for other users to run their own workflows.

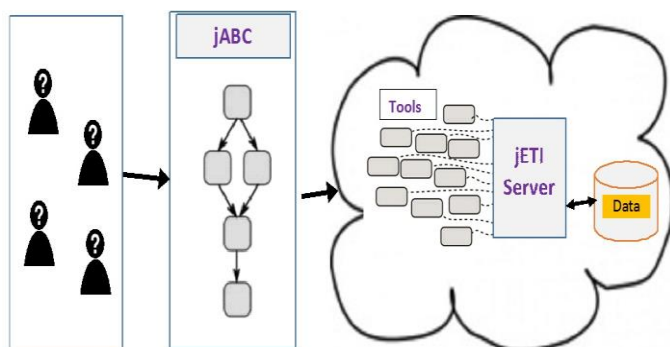


Fig.5. Interaction of users, the jABC and the jETI environment.

IV. CONCLUSION

Due to the increase of using GIS in a wide range of domains, software reuse and data sharing become more important. The service orientation paradigm has been developed to support software reuse. However, the Geospatial services have their own characteristics, such as complex processes and big data sets, that hamper the service reuse. In this light, the approach presented in this paper aims to improve the reuse of geospatial services by applying XMDD-based technologies such as jABC and jETI, and it focuses on the reuse challenge from three perspectives:

1. Performing rigorous servification by turning basic components as well as their compositions into flexibly reusable pieces of functionality,
2. enabling flexible and easy service consumption to reuse and compose services in an agile workflows which free end users from the burdens of learning programming/scripting languages and other required technologies to design and adapt workflows.
3. offering a suitable environment to handle comprehensive geospatial processing by supporting remote execution and integration of services.

In the example presented in this work, we discussed how the reuse of services used in the analysis of sea-level rise impacts is improved. The next step may be to perform a similar servification process for other (scientific) tools, extending the library with additional and alternative general and geospatial services. Moreover, a more flexible inclusion of various, heterogeneous data sources could be achieved with additional SIBs. However, to support easy and correct reuse of services also in large-scale applications, the core of our future work is going to address the semantically aware reuse of geospatial services by designing domain-specific ontologies. Once a semantics-based workflow design framework is available, the reuse of services in geospatial applications by a larger audience will become possible.

ACKNOWLEDGMENT

The authors would like to thank Steffen Kriewald, Dominik Reusser, and Markus Wrobel from the climate change and development group at the Potsdam Institute for Climate Impact Research (PIK), who provided us with the tools and data sets required for the scenario. Special thanks go to Sven Lehmann for his support with the server configuration.

This work was supported, in part, by Science Foundation Ireland grant 13/RC/2094 to Lero - the Irish Software Research Centre (www.lero.ie). Also supported, in part, by German Academic Exchange Service (DAAD) grant 2014/15 (57076385).

REFERENCES

- [1] M. J. Mineter, C. Jarvis, and S. Dowers, "From stand-alone programs towards grid-aware services and components: a case study in agricultural modelling with interpolated climate data," *Environmental Modelling & Software*, vol. 18, no. 4, 2003, pp. 379–391.
- [2] R. Lake and J. Farley, "Infrastructure for the geospatial web," in *The Geospatial Web*. plus 0.5em minus 0.4emSpringer, 2007, pp. 15–26.
- [3] L. Bernard and N. Ostländer, "Assessing climate change vulnerability in the arctic using geographic information services in spatial data infrastructures," *Climatic Change*, vol. 87, no. 1-2, 2008, pp. 263–281.
- [4] OGC Web Services Standards. [Online]. Available: <http://www.opengeospatial.org/=Opt>

- [5] T. Foerster, B. Schaeffer, J. Brauner, and S. Jirka, "Integrating OGC web processing services into geospatial mass-market applications," in *Advanced Geographic Information Systems & Web Services, 2009. GEOWS'09. International Conference on IEEE*, 2009, pp. 98–103.
- [6] V. Rautenbach, S. Coetzee, and A. Iwaniak, "Orchestrating OGC web services to produce thematic maps in a spatial information infrastructure," *Computers, Environment and Urban Systems*, vol. 37, pp. 107–120.
- [7] Y. Liu, I. Gorton, and A. Wynne, "Architecture-Based Adaptivity Support for Service Oriented Scientific Workflows," in *Service Oriented System Engineering (SOSE), 2013 IEEE 7th International Symposium on IEEE*, 2013, pp. 309–314.
- [8] A. M. Castronova, J. L. Goodall, and M. M. Elag, "Models as web services using the open geospatial consortium (OGC) web processing service (WPS) standard," *Environmental Modelling & Software*, vol. 41, pp. 72–83.
- [9] C. Granell, L. Díaz, and M. Gould, "Service-oriented applications for environmental models: Reusable geospatial services," *Environmental Modelling & Software*, vol. 25, no. 2, 2010, pp. 182–198.
- [10] G. Alonso and C. Hagen, "Geo-Opera: Workflow concepts for spatial processes," in *Advances in Spatial Databases. Springer*, 1997, pp. 238–258.
- [11] B. Ludäscher, I. Altintas, C. Berkley, D. Higgins, E. Jaeger, M. Jones, E. A. Lee, J. Tao, and Y. Zhao, "Scientific workflow management and the Kepler system," *Concurrency and Computation: Practice and Experience*, vol. 18, no. 10, 2006, pp. 1039–1065.
- [12] E. Jaeger, I. Altintas, J. Zhang, B. Ludäscher, D. Pennington, and W. Michener, "A Scientific Workflow Approach to Distributed Geospatial Data Processing using Web Services." in *SSDBM. Citeseer*, 2005, pp. 87–90.
- [13] A. Pratt, C. Peters, G. Siddeswara, B. Lee, and A. Terhorst, "Exposing the Kepler scientific workflow system as an OGC web processing service," *Proceedings of iEMSs (International Environmental Modelling and Software Society)*, vol. 2010.
- [14] N. Chen, L. Di, G. Yu, and J. Gong, "Geo-processing workflow driven wildfire hot pixel detection under sensor web environment," *Computers & Geosciences*, vol. 36, no. 3, 2010, pp. 362–372.
- [15] G. Hobona, D. Fairbairn, H. Hiden, and P. James, "Orchestration of grid-enabled geospatial web services in geoscientific workflows," *Automation Science and Engineering, IEEE Transactions on*, vol. 7, no. 2, 2010, pp. 407–411.
- [16] J. Zhang, "A practical approach to developing a web-based geospatial workflow composition and execution system," in *Proceedings of the 3rd International Conference on Computing for Geospatial Research and Applications. ACM*, 2012, p. 21.
- [17] W. Du, H. Fan, J. Li, and H. Wang, "Model-driven geospatial web service composition," *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 1, 2014, pp. 7–11.
- [18] C. Szabo, Q. Z. Sheng, T. Kroeger, Y. Zhang, and J. Yu, "Science in the cloud: Allocation and execution of data-intensive scientific workflows," *Journal of Grid Computing*, 2013, pp. 1–20.
- [19] C. Hoffa, G. Mehta, T. Freeman, E. Deelman, K. Keahey, B. Berriman, and J. Good, "On the use of cloud computing for scientific workflows," in *eScience, 2008. eScience'08. IEEE Fourth International Conference on. IEEE*, 2008, pp. 640–645.
- [20] B. Schäffer, B. Baranski, and T. Foerster, "Towards spatial data infrastructures in the clouds," in *Geospatial Thinking. Springer*, 2010, pp. 399–418.
- [21] T. Margaria and B. Steffen, "Service-Oriented: Conquering Complexity with XMDD," in *Conquering Complexity*, M. Hinchey and L. Coyle, Eds. Springer London, 2012, pp. 217–236. [Online]. Available:[http://dx.doi.org/10.1007/978-1-4471-2297-5\do5\(1\)00pt](http://dx.doi.org/10.1007/978-1-4471-2297-5_s\do5(1)00pt).
- [22] I. Crnkovic, "Component-based software engineering -new challenges in software development," *Software Focus*, vol. 2, no. 4, 2001, pp. 127–133.
- [23] T. Ravichandran, "Special issue on component-based software development," *ACM SIGMIS Database*, vol. 34, no. 4, 2003, pp. 45–46.
- [24] H. P. Breivold and M. Larsson, "Component-based and service-oriented software engineering: Key concepts and principles," in *Software Engineering and Advanced Applications, 2007. 33rd EUROMICRO Conference on. IEEE*, 2007, pp. 13–20.
- [25] J. Greenfield and K. Short, "Software factories: assembling applications with patterns, models, frameworks and tools," in *Companion of the 18th annual ACM SIGPLAN conference on Object-oriented programming, systems, languages, and applications. ACM*, 2003, pp. 16–27.
- [26] I. J. Taylor, E. Deelman, D. Gannon, M. Shields *et al.*, *Workflows for e-Science*. Springer-Verlag London Limited, 2007.
- [27] I. Foster, Y. Zhao, I. Raicu, and S. Lu, "Cloud computing and grid computing 360-degree compared," in *Grid Computing Environments Workshop, 2008. GCE'08. IEEE*, 2008, pp. 1–10.
- [28] G. Juve, E. Deelman, K. Vahi, G. Mehta, B. Berriman, B. P. Berman, and P. Maechling, "Data sharing options for scientific workflows on amazon ec2," in *Proceedings of the 2010 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis. IEEE Computer Society*, 2010, pp. 1–9.
- [29] K. Beck, M. Beedle, A. van Bennekum, A. Cockburn, W. Cunningham, M. Fowler, J. Grenning, J. Highsmith, A. Hunt, R. Jeffries, J. Kern, B. Marick, R. C. Martin, S. Mellor, K. Schwaber, J. Sutherland, and D. Thomas, "Manifesto for Agile Software Development," <http://agilemanifesto.org>, 2001, [Online; last accessed 4-March-2014].
- [30] T. Margaria and B. Steffen, "Agile IT: Thinking in user-centric models," in *Leveraging Applications of Formal Methods, Verification and Validation*, ser. Communications in Computer and Information Science, vol. 17. Springer Berlin / Heidelberg, 2009, pp. 490–502.
- [31] B. Steffen, T. Margaria, R. Nagel, S. Jörges, and C. Kubczak, "Model-driven development with the jABC," in *Hardware and Software, Verification and Testing. Springer*, 2007, pp. 92–108.
- [32] T. Margaria, C. Kubczak, M. Njoku, and B. Steffen, "Model-based design of distributed collaborative bioinformatics processes in the jABC," in *Engineering of Complex Computer Systems, 2006. ICECCS 2006. 11th IEEE International Conference on. IEEE*, 2006, pp. 8–pp.
- [33] G. Jung, T. Margaria, R. Nagel, W. Schubert, B. Steffen, and H. Voigt, "SCA and jABC: Bringing a service-oriented paradigm to web-service construction," in *Leveraging Applications of Formal Methods, Verification and Validation. Springer*, 2009, pp. 139–154.
- [34] T. Margaria, R. Nagel, and B. Steffen, "jETI: A tool for remote tool integration," in *Tools and Algorithms for the Construction and Analysis of Systems. Springer*, 2005, pp. 557–562.
- [35] A.-L. Lamprecht, *User-Level Workflow Design - A Bioinformatics Perspective*, ser. Lecture Notes in Computer Science. Springer, 2013, vol. 8311.
- [36] A.-L. Lamprecht and T. Margaria, Eds., *Process Design for Natural Scientists*, ser. CCIS. Springer, 2014, vol. 500.
- [37] S. Al-areqi, S. Kriewald, A. Lamprecht, D. Reusser, M. Wrobel, and T. Margaria, "Agile Workflows for Climate Impact Risk Assessment based on the ci: grasp Platform and the jABC Modeling Framework," in *International Environmental Modelling and Software Society (iEMSs) 7th Intl. Congress on Env. Modelling and Software (published, 2014)*.
- [38] B. Steffen, T. Margaria, and V. Braun, "The Electronic Tool Integration platform: concepts and design," *International Journal on Software Tools for Technology Transfer (STTT)*, vol. 1, no. 1, 1997, pp. 9–30.
- [39] S. Kriewald, *srtmtools: SRTM tools*, 2013, r package version 2013-00.0.1.
- [40] M. Kilibarda, "A plotGoogleMaps tutorial," 2013.
- [41] T. Margaria, "Service is in the Eyes of the Beholder," *IEEE Computer*, Nov. 2007.
- [42] A. Wickert and A.-L. Lamprecht, "jabcstats: An extensible process library for the empirical analysis of jabc workflows," in *Leveraging Applications of Formal Methods, Verification and Validation. Specialized Techniques and Applications. Springer*, 2014, pp. 449–463.

A Domain-Specific Language for Service Level Agreement Specification

Renata Vaderna, Željko Vuković, Dušan Okanović, Igor Dejanović

Faculty of Technical Sciences
University of Novi Sad
Novi Sad, Serbia
{vrenata, zeljkov, oki, igord}@uns.ac.rs

Abstract—In order to perform continuous monitoring, SLA document between interested parties has to be signed. These documents should be in machine readable format in order to automate monitoring process. On the other hand, it would be beneficial if it is human readable, too. This way, it is easier to perform configuration and maintenance of monitoring subsystem. Building up on our previous work, in this paper we present DProfLang. DProfLang is a domain specific language for defining SLAs, that are both human and machine readable.

Keywords—SLA, continuous monitoring, Domain-Specific Languages

I. INTRODUCTION

Requirements that certain software has to fulfill are usually agreed between interested parties before the start of implementation. There are two types of requirements: functional and non-functional. Ensuring that software fulfills functional requirement means that it will "do what it is expected to do." On the other hand, implementation of non-functional requirements means that the software will "do what is expected, but in a certain way." It is important to stress that while performance measurements can be performed during the development phase, it is only under production workload that we can retrieve realistic software performance data. There are often bugs that take a lot of time to manifest themselves [1], and this kind of time is not available during development. In contrast to profiling and debugging, when performing continuous monitoring we measure application performance parameters under production workload.

There is a wide array of nonfunctional requirements and metrics that can be used to quantify them. Some commonly used are response time, availability, security, robustness, memory footprint, CPU time. These parameters are usually referred to as software performance and are specified in an additional document that follows the initial agreement between the parties. This document is called Service Level Agreement (SLA). It can contain functional requirements, ways of measuring their fulfillment, referent values, ways of processing these values, and whom to contact if something goes wrong, either with the obtained values or the measuring process itself.

In our previous works [2, 3], we have described the DProf system for adaptive continuous monitoring. It is based on the Kieker monitoring framework [4], and it monitors application performance using monitoring probes. These probes are inserted into software using AspectJ or some other tool [5], and collect monitoring data, while the application is running. Adaptation of the monitoring process allows for reduction of

monitoring overhead. This is done by turning monitoring off in the call tree [6] branches that show no discrepancy between the obtained values and values specified in SLA.

SLA for the DProf system is an XML document based on the DProfSLA XML schema [2]. Since XML is a machine readable format, but not well suited for human use [20], in this paper we propose a new language - DProfLang - for monitoring goals definition. The domain specific language that we propose in this paper has the advantage of being both human and machine readable, thus allowing easier maintenance of monitoring configuration, while being well suited for monitoring automation.

The remainder of this paper is as follows. Chapter 2 shows XML schema that we currently use. In chapter 3, grammar of the new language is shown. Chapter 4 shows how to translate a document from DProfSLA format into DProfLang. Chapter 5 presents related work, while in the last section we draw conclusions and outline for the future work.

II. DPROFSLA

Root element of DProfSLA XML schema is shown in Fig. 1. It has three subelements:

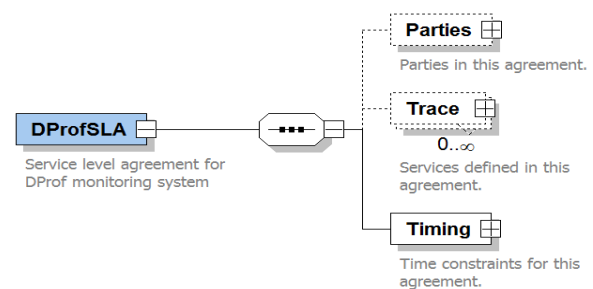


Fig. 1. Root element of DProfSLA XML schema

Parties element is simple and is used to designate interested parties and their roles in the execution of the agreement.

Timing element specifies the agreement's time constraints - the start and the end of the monitoring process, and the frequency of checkups.

Trace element (of *CallTreeNode* type - Fig. 2) is used to specify which part of the application is monitored and how the obtained data is processed. In essence, every trace element relates to one node in a call tree, i.e. a method call.

For designating call tree nodes we use attribute *name* in *CallTreeNodeType* and syntax shown in [2]. For the call tree in Fig. 3, we have the DProfSLA document from Listing 1.

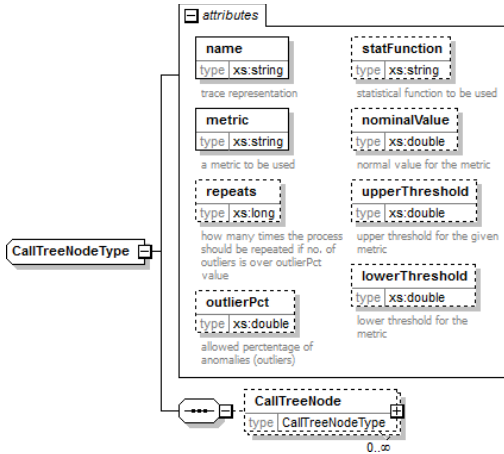


Fig. 2. Call tree node representation in DProfSLA XML schema

A node is represented with class and method name, followed by names of methods that are invoked from it. In this example, we monitor execution times, calculate averages, and compare those values to the specified upper threshold.

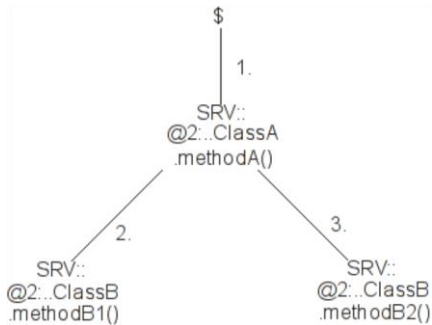


Fig. 3. An example of call tree

```
<DProfSLA>
<Parties><Provider name="Org1" />
<Consumer name="Org2" /></Parties>
<CallTreeNode metric="avgExecutionTime"
name="ClassA.methodA, [{ClassB.methodB1, []},
{ClassB.methodB2, []}]" upperThreshold="350">
<CallTreeNode metric="avgExecutionTime"
name="{ ClassB.methodB1, []}"
upperThreshold="150"/>
<CallTreeNode metric="avgExecutionTime"
name="{ ClassB.methodB2, []}"
upperThreshold="150"/>
</CallTreeNode>
</Timing>
<SamplingPeriod>600000</SamplingPeriod>
</Timing>
</DProfSLA>
```

Listing 1. DProfSLA XML for the example shown in Fig. 3.

As stated in the introductory chapter, the use of XML provides the possibility of automation of the monitoring process, since XML is machine readable. However, the use of DSL would allow human readability, while retaining machine readability.

III. DPROFLANG LANGUAGE GRAMMAR

DProfLang DSL is implemented using textX [7] meta-language and library for DSL development in Python programming language. From a single language description (grammar) textX builds a parser and a meta-model (i.e. abstract syntax) for the language.

textX grammar consists of a set of rules which define each language construct and will be translated to Python classes during Abstract Syntax Tree (AST) construction. Each rule also defines the syntax of the language element.

In Listing 2 a part of DProfLang grammar is presented. From this grammar textX will create the meta-model presented in Fig. 4. BASETYPE hierarchy is a part of the built-in textX type system.

```
DProfModel:
'SLA' name=STRING description=STRING
'parties' '{'
    parties+=Party
'}'
timing=Timing
call_node=CallNode
;

CallNode:
'cnode' name=STRING '{'
    // Inherited from parent node.
    // The root node must specify it.
    ('metric' metric=Metric)?
    // These are optional as there are
    // default values specified.
    ('repeats' repeats=INT)?
    ('outlier_percentage'
        outlier_percentage=INT)?
    ('stat_func' stat_func=StatFunc)?
    ('nominal_value' nominal_value=FLOAT)?
    ('lower_threshold' lower_threshold=FLOAT)?
    ('upper_threshold' upper_threshold=FLOAT
        nodes*=CallNode
    )
'}'
;
```

Listing 2. A part of the DProfLang grammar in textX

The DProfModel rule is the root of the meta-model. Instances of these classes have the following attributes:

- *name* – is the name of the SLA agreement,
- *description* – is an optional description given as a string,
- *parties* – is a list of the involved parties,
- *timing* – is an interval specifying when the monitoring will be applied,
- *call_node* – is the root of the call tree node hierarchy.

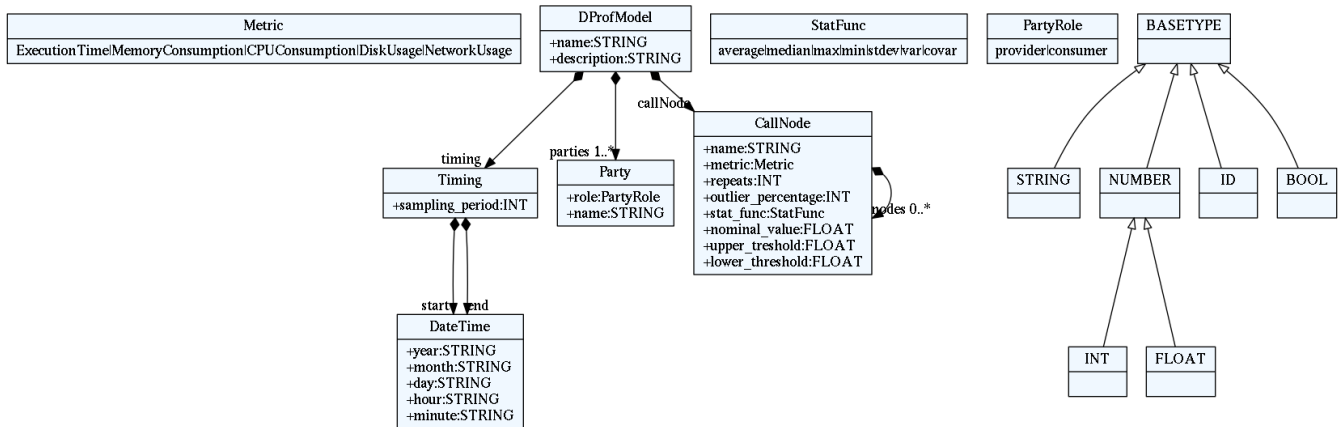


Fig. 4. DProfLang textX meta-model

CallNode rule defines a node in a call tree node hierarchy and specifies monitoring parameters such as: used metric, repeats and outlier percentage, nominal value, upper and lower threshold. This rule uses composite pattern, as each node can contain other nodes which are specified by the assignment nodes*=CallNode. textX assignment operator '*=' will match zero or more right-hand-side rules and each instance will be appended to the left-hand-side attribute.

```
SLA "Example"
parties {
    provider "Org1"
    consumer "Org2"
}
timing {
    sampling_period 600000
}
cnode "ClassA.methodA" {
    metric ExecutionTime
    stat_func Average
    upper_threshold 350
    cnode "ClassB.methodB1" {
        upper_threshold 150
    }
    cnode "ClassB.methodB2" {
        upper_threshold 150
    }
}
```

Listing 3. An example of SLA specification written in DProfLang

DProfLang meta-model instance is a Python object which is capable of parsing and instantiating DProfLang models written as DSL textual specifications.

Listing 3 shows an example of a DProfLang agreement of the DProfSLA document from Listing 1. It is obvious that the readability and comprehensibility is vastly improved with the DSL approach.

A. Transformation From DProfSLA to DProfLang

In order to integrate the new language with our previous work, we have developed two code generators. The first

generator loads DProfSLA document in the original XML format and outputs the agreement in the new DSL format. The second one does the reverse job - it parses the agreements in DProfLang format and provides XML based DProfSLA document.

For code generation, Jinja2 template engine [8] for Python has been used. A template engine is a piece of software that combines a data model with a template specification to produce a textual output. In our case data model is based on DProfLang meta-model. Two templates have been used: DProfSLA XML template and DProfLang DSL template. Instantiating data model from DProfLang DSL is supported through textX, since it automatically constructs the model from the grammar. In order to support XML we had to develop a procedure that builds data model out of DProfSLA XML.

IV. RELATED WORK

SLAs must be defined in machine-readable format to allow automatic service level management. Tebbani et al. [9] have already shown that only a few formal SLA specification languages exist. Usually, SLAs are written in some informal language, which is not acceptable for automation of the process. Therefore, authors propose Generalized Service Level Agreement language - GSLA. A GSLA document is a contract between interested parties that is designed to create a measurable common understanding of each party's role. The role is a set of rules which defines the service level expectations and obligations the party has. To specify GSLA in machine readable format, GXLA XML schema has been

proposed. Sections of GXLA documents are as follows. Schedule section contains temporal parameters of the contract. Party section models involved parties. Service package is an abstraction that is used to describe the services and previously mentioned roles. By using GXLA the service management process can be automated.

For web service SLAs, WSLA [10] can be used. It is also XML-based. Similarly to GSLA/GXLA, WSLA documents define the involved parties, metrics, measuring techniques, responsibilities, and courses of action. The authors state that every SLA language, such as WSLA, should contain 1) information regarding the agreeing parties and their roles, 2) SLA parameters and a measurement specification, as well as 3) obligations for each party.

SLAng [11] is a language for specifying SLAs based on the Meta Object Facility [12]. It can use different languages to describe constraints, e.g., utilizing OCL [13] or HUTN [14].

The WS-Agreement specification language [15] has been approved by the Open Grid Forum. It defines a language that can be used by service providers to offer services and resources, and by clients to create an agreement with that provider.

Paschke et al. [16] propose to categorize SLA metrics in order to support the design and implementation of SLAs that can be monitored and enforced automatically. Standard elements of each SLA are categorized as: technical (service descriptions, service objects, metrics, and actions), organizational (roles, monitoring parameters, reporting, and change management), and legal (legal obligations, payment, additional rights, etc.).

According to this categorization, our DProfLang documents are operation-level documents intended to be used in-house. By versatility categorization, they belong to standard agreements. As was the case with DProfSLA schema documents, we do not need all of the features of the described schemas. DProfLang is specifically designed to be used with the DProf system. Our documents provide a subset of the elements defined by GXLA or WSLA. A transformation of SLA documents between DProfLang and the mentioned schemas could, for example, be performed using appropriate generators.

Aside from XML, an SLA can be specified using domain specific languages. Most of them are AOP based, like DiSL [17], Josh [18] or Scope [19]. The problem with using AOP is that they are very platform specific. The use of a true DSL for SLA specification allows for writing of human readable documents that can be translated into instrumentation for any platform.

V. CONCLUSION

In this paper we have shown a new language for instrumentation specification. The advantage of this approach over the use of XML is that the SLA documents written with DProfLang are human readable. This allows for easier maintenance of monitoring system and better overall control over monitoring process. In contrast to the use of AOP and

AOP-like tools, our approach is platform independent. Whichever the underlying platform might be, DProfLang SLA document will be translated into instrumentation for the underlying platform.

DProfLang is implemented in textX meta-language which enables easy language grammar and meta-model modifications thus facilitating its evolution. To enable integration with our pre-existing XML based solution we have also implemented a translator from XML to the new DSL and vice versa.

Our future work will focus on development of instrumentation generators for different platforms. As DProf and Kieker use AspectJ instrumentation, our first step is to develop instrumentation generators for AspectJ. After that, our work will include generators for DiSL and .NET AOP frameworks.

ACKNOWLEDGMENT

The research presented in this paper was supported by the Ministry of Science and Technological Development of the Republic of Serbia, grant III-44010, Title: Intelligent Systems for Software Product Development and Business Support based on Models.

REFERENCES

- [1] M. Grottke, K. S. Trivedi. "Fighting Bugs: Remove, Retry, Replicate, Rejuvenate," *IEEE Computer*, v.40, n. 2, 2007, pp. 107-109.
- [2] D. Okanović, A. Van Hoorn, Z. Konjović, M. Vidaković, "SLA-Driven Adaptive Monitoring of Distributed Applications for Performance Problem Localization," *Computer Science and Information Systems*, vol. 10, no. 1, 2013, pp. 25-50.
- [3] D. Okanović, A. van Hoorn, Z. Konjović, M. Vidaković, "Towards Adaptive Monitoring of Java EE Applications", *Proceedings of the 5th International Conference on Information Technology - ICIT*. Amman, Jordan, 2011, CD.
- [4] A. van Hoorn, W. Hasselbring, J. Waller, "Kieker: A Framework for Application Performance Monitoring and Dynamic Software Analysis," *Proceedings of the 3rd ACM/SPEC International Conference on Performance Engineering (ICPE 2012)*, Boston, USA, 2012, pp. 247-248.
- [5] D. Okanović, M. Vidaković, "Evaluation of Alternative Instrumentation Frameworks," *Symposium on Software Performance: Joint Descartes/Kieker/Palladio Days*, Stuttgart, Germany, 2014, pp. 83-90.
- [6] W. Binder, J. Hulaas, P. Moret, "Advanced Java Bytecode Instrumentation," *5th International Symposium on Principles and Practice of Programming in Java*, Lisboa, Portugal, 2007, p. 135-144.
- [7] textX [Online] <https://github.com/igordejjanovic/textX> (January 2015)
- [8] Jinja2 [Online] <http://jinja.pocoo.org/docs/dev/> (January 2015)
- [9] B. Tebbani, I. Aib, "GXLA a Language for the Specification of Service Level Agreements," *Lecture Notes in Computer Science*, v. 4195. Springer-Verlag, Berlin Heidelberg New York, 2006, p. 201-214.
- [10] A. Keller, H. Ludwig, "The WSLA Framework: Specifying and Monitoring Service Level Agreements for Web Services," *Journal of Network and Systems Management*, vol. 11, no. 1, 2003, pp. 57-81.
- [11] D. Lamanna, J. Skene, W. Emmerich, "SLAng: A Language for Defining Service Level Agreements," *Proceedings of the 9th IEEE Workshop on Future Trends of Distributed Computer Systems (FTDCS '03)*, IEEE Computer Society, San Juan, Puerto Rico, 2003, pp. 100-107.
- [12] Meta Object Facility (MOF) 2.0 Core Specification. OMG. [Online] Available: <http://www.omg.org/spec/MOF/2.0> (current September 2011)

- [13] Object Constraint Language (OCL) 2.0. OMG. [Online] Available: <http://www.omg.org/spec/MOF/2.0> (January 2015)
- [14] Human Usable Textual Notation (HUTN) Specification. OMG. [Online] Available: <http://www.omg.org/spec/HUTN/index.htm> (January 2015)
- [15] N. Oldham, K. Verma, A. Sheth, F. Hakimpour, "Semantic WS-agreement partner selection," 15th International Conference on World Wide Web. ACM, Edinburgh, Scotland, UK, 2006, pp. 697-706.
- [16] A. Paschke, E. Schnappinger-Gerull, "A Categorization Scheme for SLA Metrics," Multi-Conference Information Systems (MKWI 2006), Passau, Germany, 2006, pp. 25-40.
- [17] L. Marek, A. Villazón, Y. Zheng, D. Ansaloni, W. Binder, Z. Qi, "DiSL: a Domain Specific Language for Bytecode Instrumentation," 11th Annual International Conference on Aspect-Oriented Software Development (AOSD '12), 2012, pp. 239-250.
- [18] S. Chiba, K. Nakagawa, "Josh: an Open AspectJ-Like Language, ". AOSD'04, ACM, 2004, pp. 102-111.
- [19] T. Aotani, H. Masuhara, "Scope: an AspectJ Compiler for Supporting User-Defined Analysis-Based Pointcuts," AOSD'07, ACM, 2007, pp. 161-172.
- [20] T. Parr, "Humans should not have to grok XML; Answers to the question 'When shouldn't you use XML?'," IBM DeveloperWorks, 2001

Anatomy of the Tree Based Strategy for High Strength Interaction Testing

Mohammad F. J. Klaib

Computer Science Department
College of Computer Sciences and Engineering
Taibah University
Madina, Kingdom of Saudi Arabia
Email: mklaib@taibahu.edu.sa
mom_klaib@yahoo.com

Abstract—The amount of resources consumed for a complete and exhaustive testing becomes unreasonable and unaffordable. While it is vital to assure the quality and the reliability of any system, it is impossible to do an exhaustive testing due to the huge number of possible combinations. To bring a balance between exhaustive testing and lack of testing combinatorial interactions testing has been adopted. Although it is stated in literature that a complete pairwise interaction testing ensures the detection of 50–97 percent of faults, it is not sufficient to stop with pairwise testing alone for highly interactive systems. Therefore, there is a need to extend the level of testing for a general multi way combinatorial interactions testing. This paper enhanced the previous strategies “A tree based strategy for test data generation and cost calculation” and “3-way interaction testing using the tree strategy” to support a general multi-way combinatorial interaction testing involving uniform and non uniform parametric values. In this strategy, two algorithms have been adopted; a tree construction algorithm which constructs the possible test cases and an iterative cost calculation algorithm that constructs efficient multi-way test suites which cover all parameter interactions between input components. Both algorithms are presented in details.

Keywords— Software testing, Hardware testing, Multi-way testing

I. INTRODUCTION

Testing [1] is an activity aims to evaluate the attributes or capabilities of software or hardware products, and determines if the products have met their requirements. Testing in general is a very important phase of the development cycle for both software and hardware products [2-5]. Testing helps to reveal the hidden problems in the product, which otherwise goes unnoticed providing a false sense of well-being. It is said to cover 40 to 50 percent of the development cost and resources [6,7]. Although important to quality and widely deployed by programmers and testers, testing still remains an art. A good set of test data is one that has a high chance of uncovering previously unknown errors at a faster pace. For a successful test run of a system, we need to construct a good set of test data covering all interactions among system components [34-39].

Failures of hardware and software systems are often caused due to unexpected interactions among system components. The failure of any system may be catastrophic that we may lose very important data or fortunes or sometimes even lives [7,8]. The main reason for failure is the lack of proper testing. A complete test requires testing all possible combinations of interactions, which can be exorbitant even for medium sized projects due to the huge number of combinations (Combinatorial explosion problem).

Combinatorial Explosion – All products are built with basic elements which interact with one another by means of predefined combination rules. As the number of classes of elements increases, the number of interactions between the elements also increases exponentially [9-11] which leads to the

problem of combinatorial explosion. Thus, combinatorial explosion [21,22] occurs when a huge number of possible combinations are produced by increasing the number of entities or elements, which have to interact with one another for successful functioning of a product.

To gain a better understanding of this problem, we consider a simple example of testing a 16-1 multiplexer. A multiplexer or mux is a device that selects one of many analog or digital input signals and forwards the selected input to a single output line. A multiplexer of 2^n inputs has n select lines, which are used to select which input line will be directed the output. Fig. 1 below shows a black box of 16-1 multiplexer. In order to exhaustively test such a multiplexer, there are 2^{20} (i.e. 1048576) combinations of tests that needs to be performed. If the time required for one test to be executed is 5 minutes, then it would take nearly 10 years for a complete test to be done. Thus, the amount of resources consumed for a complete and exhaustive testing of the system becomes unreasonable and unaffordable [12,13]. While it is vital to assure the quality and the reliability of the system, it is impossible to do an exhaustive testing due to the combinatorial explosion problem. Therefore, it is very clear that combinatorial explosion is a serious and critical issue that all software and hardware testers face.

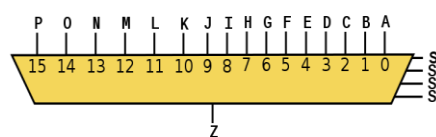


Fig 1 16-1 Multiplexer

Thus, to bring a balance between exhaustive testing and lack of testing, combinatorial interaction testing [14-16] has demonstrated to be an effective technique to achieve reduction of test suite size, thus relieving the problem of combinatorial explosion. Combinatorial interaction testing, samples the systems input space and produces a set of factor-value bindings that typically cover all possible pairs or multi-way combinations of factor-values, thereby achieving a high-degree of coverage and fault detection [17].

Testing all pairwise (2-way) interactions between input components ensure the detection of 50 – 97 percent of faults [17-19], [24-30]. Although using pairwise testing gives a good percentage of reduction in fault coverage, empirical studies show that pairwise testing is not sufficient enough for highly interactive systems [23]. Therefore, there is a need to extend the level of testing to support higher multi-way combinatorial interactions, which requires every combination of any T parameter values to be covered by at least one test case, where T is referred to as the strength of coverage. Constructing a minimum test set for multi-way combinatorial interaction is still a NP complete problem [19,20] and there is no strategy that can claim that it has the best generated test suite size for all cases and systems.

Therefore, based on the above argument, this new work extends our previous strategy “A Tree Based Strategy for Test Data Generation and Cost Calculation” to go beyond pairwise combinatorial interaction testing involving uniform and non-uniform parametric values. We have two algorithms, a tree generation algorithm which generates the test cases and an iterative cost calculation algorithm which enables a minimum multi-way test data generation. The remainder of this paper is organized as follows. Section 2 presents the related work. In Section 3, the proposed tree generation and the iterative cost calculation strategies are illustrated and the correctness of both strategies have been proved with an example. Section 4 provides the conclusion.

II. RELATED WORK

There are a number of strategies proposed in literature for test suite generation of combinatorial interaction testing. Most of these strategies work only for pairwise combinatorial software interaction testing and a few others have been extended to work for T-way testing. Combinatorial interaction testing strategies could be broadly classified into two types [31] based on the approach that is used to solve the problem. They are:

- Algebraic strategies
- Computational strategies

Algebraic approaches have pre-defined rules to compute test suites directly from mathematical functions [31]. On a contrary, computational approaches use search technique to search the combinations space to generate the test cases until all T-way combinations of interactions to be covered. A number of

researches have worked in this field and have adopted either the computational or algebraic approaches.

The classification of strategies used for combinatorial software testing has been further extended by Grindal et al. [19, 20] into three main categories based on the randomness of the implemented solution. They are:

- Deterministic strategies
- Non-deterministic strategies
- Compound strategies

A deterministic strategy is one which has the property that it produces the same test suite for every execution. A non-deterministic strategy on the other hand has the property that for every execution, there is always a randomly generated combination suite to cover all the required T-way combinations. In a compound strategy two or more combination of strategies are used together.

The Automatic Efficient Test Generator or AETG [9, 14] and its variant mAETG [31] employ the computational approach. This approach uses ‘Greedy technique’ to construct test cases based on the criteria that every test case covers as many uncovered combinations as possible. The AETG uses a random search algorithm and hence the test cases are generated in a highly non-deterministic fashion [22]. Other variants of AETG use the Genetic Algorithm, Ant Colony Algorithm [20].

In Genetic algorithm [20] an initial population of individuals (test cases) are created and then the fitness of the created individuals is calculated. Then the individual selection methods are applied to discard the unfit individuals. The genetic operators such as crossover and mutation are applied to the selected individuals and this continues until we evolve a set of best individuals or the stopping criteria is attained. Thus this approach follows a non deterministic methodology similar to the Ant Colony Algorithm [20] in which each path from start to end point is associated with a candidate solution. The candidate solution is the amount of pheromone deposited on each edge of the path followed by an ant, when it reaches the end point. When an ant has to choose among the different edges, it would choose the edge with a large amount of pheromone with higher probability thus leading to better results. In some cases, these algorithms give optimal solution than original AETG.

The In-Parameter-Order [25] or IPO Strategy for pairwise testing starts constructing the test cases by considering the first two parameters, then uses a horizontal growth strategy which extends to cover the third, fourth, fifth etc. until all the parameters are considered. Further it adopts a vertical growth strategy which helps in covering all the pairs that are not covered, until all the pairs in the covering array are covered. Thus this approach generates the test cases in a deterministic fashion. Covering one parameter at a time gives a lower order of complexity to this strategy than AETG. The IPOG [8, 16] strategy extends IPO, so that IPOG can generate test suite supporting T-way combinatorial interactions. The IRPS Strategy [33] uses the computational approach and so generates all pairs and stores them in a linked list and then searches the list to arrive at the best set of test cases in a deterministic fashion.

The G2Way [13] uses a computational and deterministic strategy. It adopts a backtracking strategy to generate the test cases. The main algorithms that form the G2Way strategy consist of the parser algorithm, the 2-way combination generation algorithm, the backtracking algorithm, and the executor algorithm. The parser algorithm will load the parameter and values to be used by the 2-way combination generation

adds it to the list of restrictions. Thus it uses a computational and deterministic approach for test suite generation.

WHITCH is IBM's Intelligent Test Case Handler. With the given coverage properties it uses combinatorial algorithms to construct test suites over large parameter spaces. TVG [30] is a free tool that is built based on model based techniques. It combines both behavior and data modelling techniques. The

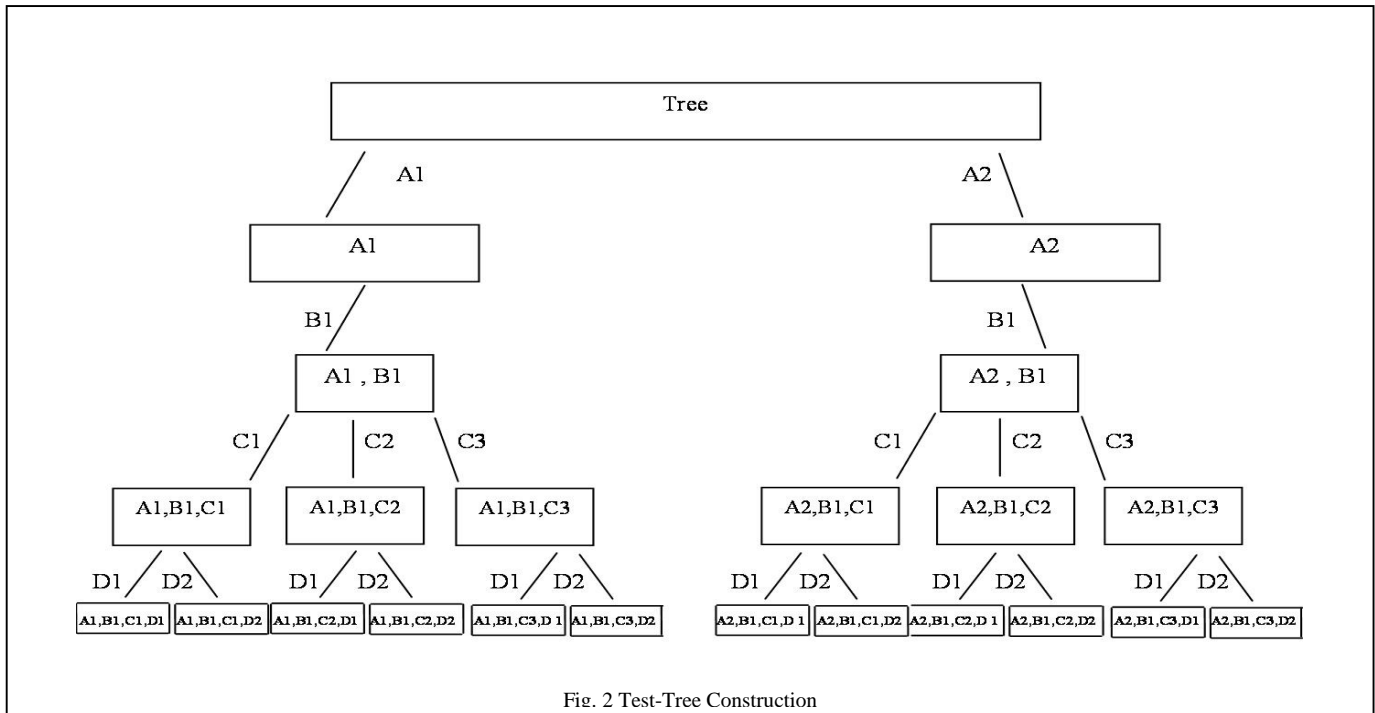


Fig. 2 Test-Tree Construction

algorithm which generates the 2-way covering array. Exploiting the result generated by the combination generation algorithm, the backtracking algorithm generates the 2-way test sets in two phases. In the first phase, the sets generated by the combination generation algorithm are merged together to form complete test suites. In the second phase, all the test sets in the generated test suite are checked to ensure that all the combinations in the covering array are covered. GTWay adopts the same strategies as that of G2Way but generates test suites for general and high T-way combinatorial interaction strengths.

The TConfig [28] uses a deterministic approach to construct test suites for T-way testing. It uses a recursive algorithm for pairwise interaction testing and a version of IPO for T-way testing. TConfig was mainly developed for pairwise interaction test suite generation by applying the theory of orthogonal Latin squares from balanced statistical experiments. Jenny [29] is a tool similar to AETG, which first covers single features (one way interaction), then pairs (2-way interaction) of features, then triples (3-way interaction), and so forth up to the n-tuples requested by the user. During each pass it checks whether the existing tests cover all tuples, and if not, make a list of uncovered tuples and add more tests until all tuples are covered. It tries to find test cases that obey the restrictions and cover a lot of new tuples. Any tuple that it can't cover no matter how hard it tries without disobeying some restriction, it says it can't cover it, and

behavior modelling allows the testers to capture important high level scenarios to test. A data model is then created at a level of sophistication according to the importance of each test scenario.

Other researchers have adopted heuristic search techniques [32] such as Hill climbing, Simulated Annealing, Tabu search, Great Flood etc. All of these search strategies have the same goal as to maximize the number of tuples covered in a test. It initially uses greedy algorithm to choose each test and then it is modified using local search. These Heuristic search techniques predict the known test set in advance in contrast to AETG and IPO which builds the test set from the scratch. However, there is no guarantee that the test set produced by Heuristic techniques are the most optimum. The AETG or IPO takes longer time to complete when compared to the Heuristic techniques. Although some work has been done in the past by researchers, test suite generation for combinatorial interaction testing still remains a research area and NP complete problem that needs exploration.

III. THE PROPOSED STRATEGY

The proposed strategy starts by constructing the test-tree based on the input parameters and values. Then it constructs the covering array, which includes all possible multi-way combinations of input variables. In order to construct the test-

tree it considers one parameter at a time until all the values of all the parameters are considered. To illustrate the concept consider a simple system with parameters and values as shown below:

- Parameter A has two values A1 and A2
- Parameter B has one value B1
- Parameter C has three values C1, C2 and C3

covered by any test case for the given set of parameters and values. Then it iterates to calculate the cost of each and every leaf node which represents the test cases, in a sequential order. The cost of any leaf node or test case is equal to the number of pairs that it covers in the covering array.

TABLE 1. PAIRWISE INTERACTION COVERING ARRAY

Test Case No.	Test Case	Iteration	Max Weight	Covered pairs
T1	A1,B1,C1,D1	1	6	[A1,B1][A1,C1][A1,D1] [B1,C1][B1,D1][C1,D1]
T10	A2,B1,C2,D2	1	6	[A2,B1][A2,C2][A2,D2] [B1,C2][B1,D2][C2,D2]
T6	A1,B1,C3,D2	2	4	[A1,C3][A1,D2] [B1,C3][C3,D2]
T11	A2,B1,C3,D1	3	3	[A2,C3] [A2,D1] [C3,D1]
T3	A1,B1,C2,D1	4	2	[A1,C2] [C2,D1]
T8	A2,B1,C1,D2	4	2	[A2,C1] [C1,D2]

Parameter D has two values D1 and D2

A with B	A with C	A with D	B with C	B with D	C with D
A1,B1	A1,C1	A1, D1	B1,C1	B1, D1	C1, D1
A2,B1	A1,C2	A1, D2	B1,C2	B1, D2	C1, D2
	A1,C3	A2, D1	B1,C3		C2, D1
	A2,C1	A2, D2			C2, D2
	A2,C2				C3, D1
	A2,C3				C3, D2

We have given the illustration for minimum test suite construction of 2-way and 3-way combinatorial interactions using our algorithm, for the example in Fig. 2 above, which depicts the system mentioned. The tree generation algorithm starts by constructing the test-tree. It uses the values of the first parameter to construct the base branches of the test-tree. Then it uses all the values of the second parameter for the next level and then the third and so on. Thus, the tree is constructed iteratively until all the parameters are considered. As a result we get all possible test cases generated for all the parameters by considering all its values.

Fig. 2 above shows how the test-tree would be constructed. The test cases generated by the test-tree are stored in the list T in a sequential order i.e. T1(A1,B1,C1,D1), T2(A1,B1,C1,D2), T3(A1,B1,C2,D1), T4(A1,B1,C2,D2), T5(A1,B1,C3,D1), T6(A1,B1,C3,D2), T7(A2,B1,C1,D1), T8(A2,B1,C1,D2), T9(A2,B1,C2,D1), T10(A2,B1,C2,D2), T11(A2,B1,C3,D1) and T12 (A2,B1,C3,D2).

The algorithm then constructs the covering array, for all possible 2-way combinations of input variables. Table 1 shows the covering array for pairwise combinations i.e. [A & B], [A & C], [A & D], [B & C], [B & D] and [C & D]. The covering array for the above example has 23 pairwise interactions which have to be covered by any test suite generated, to enable a complete pairwise interaction testing of the system.

Once the test-tree construction is over we have all the test cases generated. The next step generates the covering array, after which the cost array corresponding to the number of test cases (or leaf nodes) is created and initialised to some high value. Then, the cost calculation begins. The algorithm first calculates the maximum cost or maximum number of pairs that can be

Once it reaches a leaf node with the maximum cost, it deletes this leaf node from the list of leaf nodes generated by the test-tree i.e. T and includes this node or test case into the new list T' which holds all the test cases that are to be included in the test suite. It also deletes all the pairs that this test case has covered from the covering array. In the above example, when the first iteration begins, the first leaf node (A1,B1,C1,D1) will be deleted from T and added to T' since it has a cost equals the maximum cost 6 and the six pairs covered by it ([A1,B1], [A1,C1], [A1,D1], [B1,C1], [B1,D1] and [C1,D1]) will be deleted from the covering array. Thus, the first leaf node (or test case) generated by the test-tree will always have the maximum cost and is said to be included in T' by default for any system.

The algorithm will then continue calculating the cost of all the leaf nodes in a sequential order and includes the ones having the maximum cost. If all the pairs in the covering array are covered then the algorithm stops else it goes to the second iteration. In the second iteration, the maximum cost value (Wmax) will be decreased by one and the next set of best test cases (i.e. test cases that can cover the new Wmax number of the covering array. Thus, the algorithm continues until all the pairs in the covering array are covered. For the above example all the test cases which are included in the test suite are identified in four iterations and there are six such test cases.

Table 2 shows how the cost calculation algorithm works iteratively to generate the test suite. Table 2 also shows the order in which the various test cases are actually included in the test suite. Thus, all the 23 pairs generated for covering all pairwise

interactions as shown in Table 1, has been covered by the test cases generated by our algorithm as shown in the fifth column of the Table2. Thus this proves the correctness of our strategy in generating pairwise test suite. We have also proved that our algorithm is efficient in achieving a good reduction in the number of test cases. The exhaustive number of test cases is 12 and we have generated 6 test cases which covers all the pairs in the covering array thus achieving a 50% reduction in this case.

i.e. [A, B, C], [A, B, D], [A, C, D] and [B, C, D], for the example in Fig. 2. The covering array for the above example has 28 combinations of 3-ways interactions which have to be covered in the final test suite. Table 4 shows how the cost calculation algorithm works iteratively to generate the test suite. It also shows the order in which the various test cases are actually included in the test suite. All the 28 3-way combinations in the covering array have been covered by our algorithm. Thus, the correctness of our strategy for 3-way interaction coverage has been proved.

A. Test—Tree Construction Algorithm

The tree generation strategy thus provides the following advantages:

TABLE3. 3-WAY INTERACTION COVERING ARRAY

A, B, C	A, B, D	A, C, D	B, C, D
A1, B1, C1	A1, B1, D1	A1, C1, D1	B1, C1, D1
A1, B1, C2	A1, B1, D2	A1, C1, D2	B1, C1, D2
A1, B1, C3	A2, B1, D1	A1, C2, D1	B1, C2, D1
A2, B1, C1	A2, B1, D2	A1, C2, D2	B1, C2, D2
A2, B1, C2		A1, C3, D1	B1, C3, D1
A2, B1, C3		A1, C3, D2	B1, C3, D2
		A2, C1, D1	
		A2, C1, D2	
		A2, C2, D1	
		A2, C2, D2	
		A2, C3, D1	
		A2, C3, D2	

TABLE4. GENERATED TEST SUITE FOR 3-WAY COMBINATORIAL INTERACTION

Test Case No.	Test Case	Iteration	Max Weight	Covered pairs
T1	A1,B1,C1,D1	1	4	[A1,B1,C1][A1,B1,D1][A1,C1,D1][B1,C1,D1]
T4	A1,B1,C2,D2	1	4	[A1,B1,C2][A1,B1,D2][A1,C2,D2][B1,C2,D2]
T8	A2,B1,C1,D2	1	4	[A2,B1,C1][A2,B1,D2][A2,C1,D2][B1,C1,D2]
T9	A2,B1,C2,D1	1	4	[A2,B1,C2][A2,B1,D1][A2,C2,D1][B1,C2,D1]
T5	A1,B1,C3,D1	2	3	[A1,B1,C3][A1,C3,D1][B1,C3,D1]
T12	A2,B1,C3,D2	2	3	[A2,B1,C3][A2,C3,D2][B1,C3,D2]
T2	A1,B1,C1,D2	3	1	[A1,C1,D2]
T3	A1,B1,C2,D1	3	1	[A1,C2,D1]
T6	A1,B1,C3,D2	3	1	[A1,C3,D2]
T7	A2,B1,C1,D1	3	1	[A2,C1,D1]
T10	A2,B1,C2,D2	3	1	[A2,C2,D2]
T11	A2,B1,C3,D1	3	1	[A2,C3,D1]

After the pairwise test suite is generated, we move to the next iteration where we generate the test suite for 3-way combinatorial interactions and so on and so forth until (n-1) way combinatorial interaction test suite is generated. To illustrate the 3-way test suite generation, again the whole process starts by constructing the 3-way covering array and the iterative cost calculation of the test cases in a sequential order as explained before. Table 3 shows the covering array for 3-way combination

- A systematic method whereby all possible test cases are generated in order.
- The above procedure works well for both parameters with uniform and non-uniform values. Therefore all parameters can have different or same values as any real time system to be tested would have.

- The algorithm generates only a set of leaf nodes at every stage, although it appears as if the entire tree gets generated in order to minimise the space requirements. Therefore we only have a list of leaf nodes (or test cases) when the algorithm ends.

The example tree shown in Fig. 2 explains how the test cases are constructed manually. In reality we may need only the leaf nodes and all the intermediate nodes are not used. Therefore in order to increase the efficiency and to minimise memory allocation, we have constructed the tree shown in Fig. 2 using the proposed tree generation algorithm, which constructs the tree by minimising the number of nodes and by giving importance to only the leaf nodes at every stage.

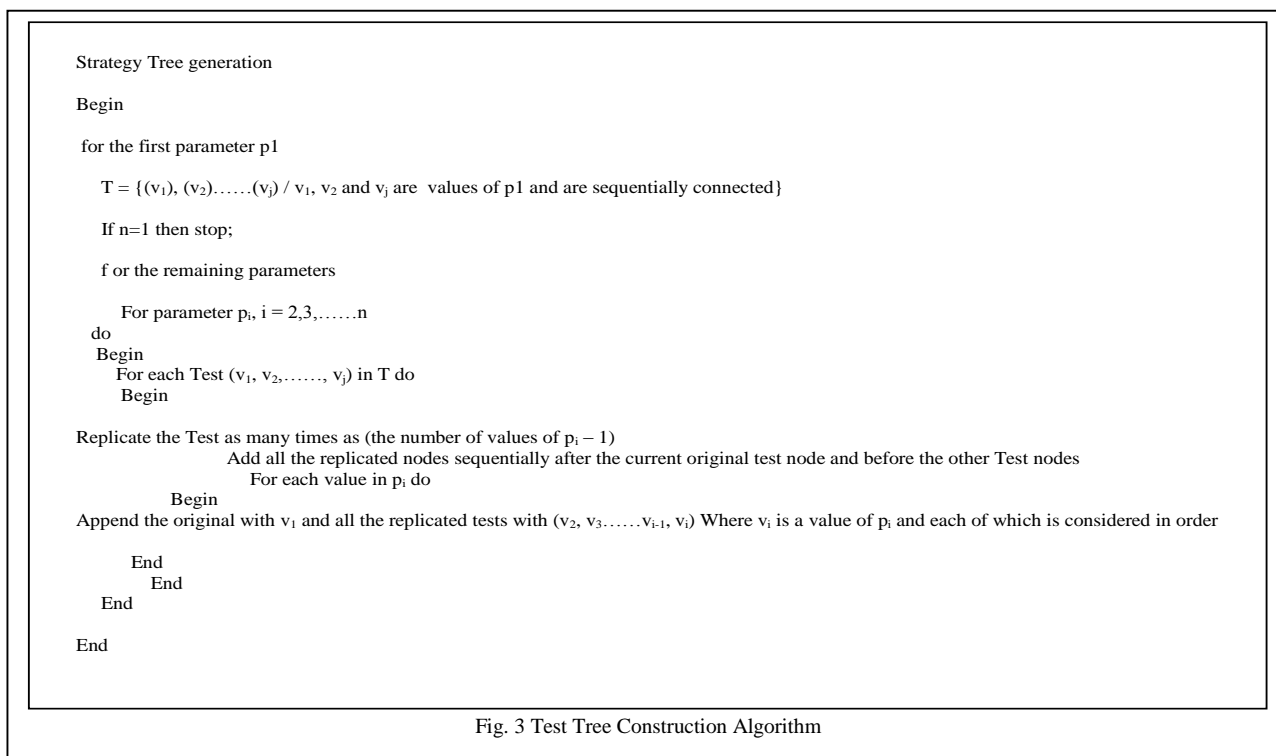
$$N_{soln} = E_{soln} * n \quad (1)$$

Assume there are 6 leaf nodes in existing set (i.e. $E_{soln}=6$), and the next parameter to be considered has 2 values (i.e. $n=2$). Then based on Equation 1 the new list of nodes will have 12 new leaf nodes as a result (i.e. $N_{soln}=12$). Therefore at every stage of tree construction, the algorithm considers each and every existing leaf node separately and calculates the number of times this particular node needs to be replicated in order to get the new set of leaf nodes with the formulae given below:

$$\text{The number of values of } p_i - 1 \quad (2)$$

Where p_i – is the i th parameter under consideration for constructing the new set of leaf nodes and $i=1,2,\dots,N$ – the number of parameters.

In Figure 2, consider the existing nodes (A1, B1) and (A2, B1). To construct the next level of nodes the parameter under consideration is C which has values C1, C2 and C3. Therefore, the node (A1, B1) needs to be replicated twice. Now we will



Therefore, at each stage or iteration we look at the leaf nodes of the tree and generate the next level nodes by considering all the values of the current parameter to generate the new set of nodes. The new set of leaf nodes from an already existing set is calculated using a replication strategy. If the existing set of leaf nodes is E_{soln} , new set of leaf nodes is N_{soln} and the number of values of the parameter under consideration is n . Then,

have three (A1, B1) nodes to which C1 is added to the first, C2 is added to the second and C3 is added to the third and then the two replicated nodes are included in the list of leaf nodes after the original node and before the node (A2, B1). The same is done to (A2, B1). It is replicated twice and hence we have three of it (one original and two replicated nodes). Now C1 is added to the first (original node), C2 is added to the second (replicated node) and C3 is added to the third (replicated node). Thus we have (A2, B1, C1), (A2, B1, C2) and (A2, B1, C3). The same process is

done to construct the test-tree until all the parameters are considered. Thus, once the list of leaf nodes is generated by considering all the values of all the parameters, we proceed to the next strategy of iterative cost calculation to construct the test suite.

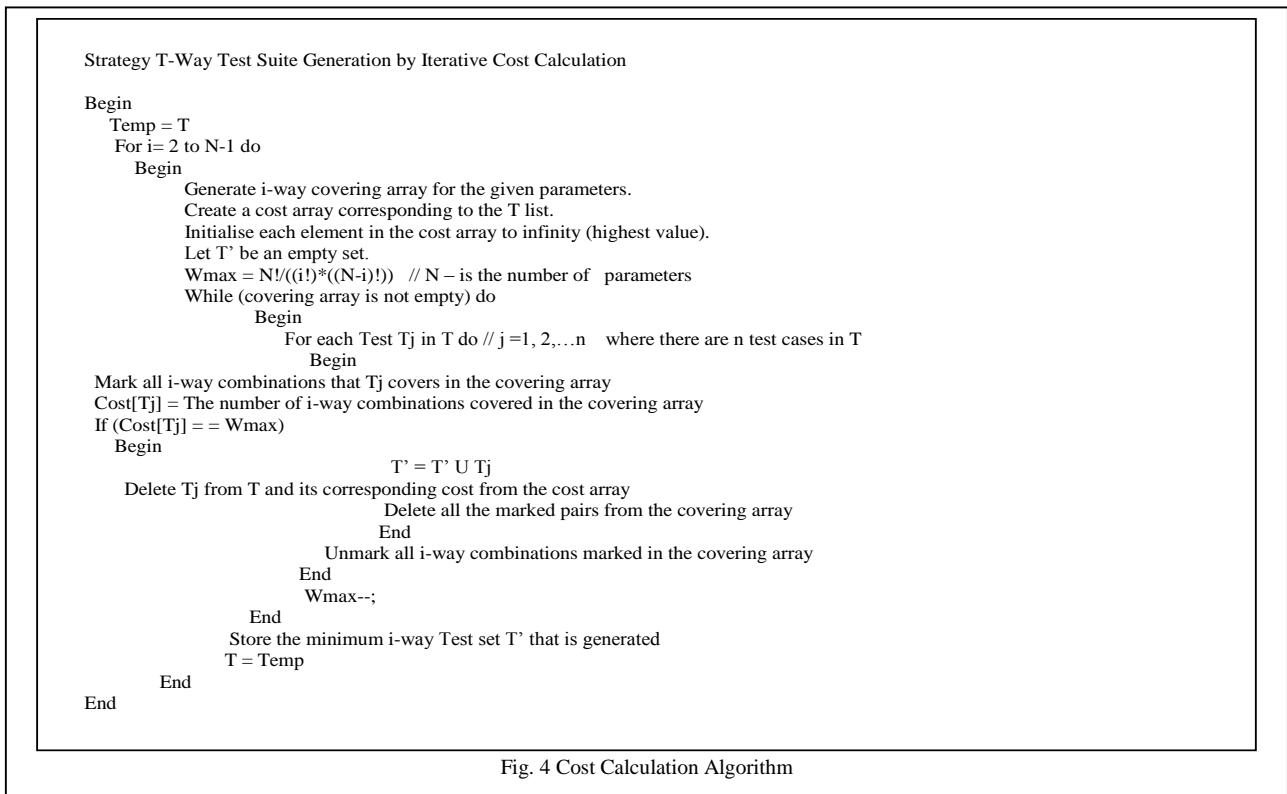
B. Iterative Cost Calculation Strategy

In Figure 4, the outer loop iterates N-2 times through the list of test cases to generate N-2 test suites, i.e. 2-way, 3-way, 4-way etc. until (N-1) way. For every T-way (i.e. 2-way, 3-way, 4-way etc. until (N-1) way) test suite to be generated, the inner loop of the algorithm iterates until all the combinations of the corresponding T-way covering array are covered. During each iteration, all the test cases with the maximum cost (Wmax) will be included in the test suite. Thus the algorithm guarantees identifying a minimum set of test cases for parameters with uniform and non-uniform values.

combinatorial interactions between input components. The correctness of the proposed strategy has been proved in Tables 2 and 4.

REFERENCES

- [1] C. Kaner, "Exploratory Testing," in Proc. of the Quality Assurance Institute Worldwide Annual Software Testing Conference, Orlando, FL, 2006.
- [2] R. Bryce, C. J. Colbourn, and M. B. Cohen, "A Framework of Greedy Methods for Constructing Interaction Tests," in Proc. of the 27th International Conference on Software Engineering, St. Louis, MO, USA, 2005, pp. 146-155.
- [3] F. F. Tsui and O. Karam, Essentials of Software Engineering. Massachusetts, USA: Jones and Bartlett Publishers, 2007.
- [4] L. G. Hernandez, J. T Jimenez, N. R Valdez, J. B. Rios, "A Post-optimization Strategy for Combinatorial Testing: Test Suite Reduction through the Identification of Wild Cards and Merge of Rows", in Advances in Computational Intelligence Lecture Notes in Computer Science vol. 7630, 2013pp 127-138.
- [5] J. Zhou, J. Liu, J. Wu, and G. Zhong, "A Latent Implementation Error



IV. CONCLUSION

In this paper we extend and improve our previous strategy, “A Tree Based Strategy for Test Data Generation and Cost Calculation” to support higher testing strength interactions. The proposed strategy is based on two algorithms. A tree construction algorithm which constructs the possible test cases and an iterative cost calculation algorithm that constructs efficient multi-way test suites which cover all possible

Detection Method for Software Validation", Journal of Applied Mathematics. 2013pp 1-10.

- [6] B. Beizer, Software Testing Techniques, 2 ed. NY: Thomson Computer Press, 1990.
- [7] M.F.J. Klaib, K.Z. Zamli, N.A.M. Isa, M.I. Younis, "G2Way A Backtracking Strategy for Pairwise Test Data Generation" Software Engineering Conference, 2008. APSEC 08, 2008, pp. 463 – 470.
- [8] K. Z. Zamli, M. F.J. Klaib, M. I. Younis, N. A. M. Isa, R. Abdullah, "Design and implementation of a t-way test data generation strategy with

- automated execution tool support". *Journal of Information Sciences*. vol 181(9), 2011, pp. 1741-1758.
- [9] S. Khatun, K. F. Rabbi, C. Y. Yaakub and M. F. J. Klaib, "A Random Search Based Effective Algorithm for Pairwise Test Data Generation" in *proc. of IEEE International Conference on Electrical Control and Computer Engineering 2011*, Kuantan, Malaysia, 2011, pp 293 - 297.
- [10] X. Qu and M. B. Cohen. "A study in prioritization for higher strength combinatorial testing". *The 2nd International Workshop on Combinatorial Testing*, 2013.
- [11] T. Nanba, T. Tsuchiya, and T. Kikuno. "Using satisfiability solving for pairwise testing in the presence of constraints". *IEICE Transactions*, vol 95(9), 2012, pp.1501–1505.
- [12] D. K. R. Chaudhuri and T. Zhu, "A Recursive Method for Construction of Designs," *Discrete Mathematics - Elsevier*, vol. 106, 1992, pp. 399-406.
- [13] M. F. J. Klaib, K. Z. Zamli, N. A. M. Isa, M. I. Younis, and R. Abdullah, "G2Way – A Backtracking Strategy for Pairwise Test Data Generation," in *Proc. of the 15th IEEE Asia-Pacific Software Engineering Conf.*, Beijing, China, 2008, pp. 463-470.
- [14] R. C. Bryce, S. Sampath, J. B. Pedersen, and S. Manchester. "Test suite prioritization by cost-based combinatorial interaction coverage". *International Journal of Systems Assurance Engineering and Management*, vol 2(2) , 2011, pp.126–134.
- [15] J. Petke, S. Yoo, M. B. Cohen, M. Harman, "Efficiency and early fault detection with lower and higher strength combinatorial interaction testing", in *Proc.ESEC/FSE 2013 Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering*, pp. 26-36, USA, 2013.
- [16] S. Varshney, M. Mehrotra. Search based software test data generation for structural testing: a perspective, in *ACM SIGSOFT Software Engineering Notes archive*, vol 38 (4), NY, USA , 2013, pp. 1-6.
- [17] M. B. Cohen, J. Snyder, and G. Roethermel, "Testing Across Configurations: Implications for Combinatorial Testing," in *Proc. of the 2nd Workshop on Advances in Model Based Software Testing*, Raleigh, North Carolina, USA, 2006, pp. 1-9.
- [18] C. J. Colbourn, M. B. Cohen, and R. C. Turban, "A Deterministic Density Algorithm for Pairwise Interaction Coverage," in *Proc. of the IASTED Intl. Conference on Software Engineerin*, Innsbruck, Austria, 2004, pp. 345-352.
- [19] K. C. Tai and Y. Lei, "A Test Generation Strategy for Pairwise Testing," *IEEE Transactions on Software Engineering*, vol. 28, 2002, pp. 109-111.
- [20] T. Shiba, T. Tsuchiya, and T. Kikuno, "Using Artificial Life Techniques to Generate Test Cases for Combinatorial Testing," in *Proc. of the 28th Annual Intl. Computer Software and Applications Conf. (COMPSAC'04)*, Hong Kong, 2004, pp. 72-77.
- [21] Mats Grindal, "Handling Combinatorial Explosion in Software Testing", *Linkoping Studies in Science and Technology*, Dissertation No. 1073, Sweden, 2007
- [22] K. Z. Zamli, N. A. M. Isa, M. F. J. Klaib, Z. H. C. Soh and C. Z. Zulkifli, "On Combinatorial Explosion Problem for Software Configuration Testing," in *Proc. of the International Robotics, Vision, Information and Signal Processing Conference (ROVISP2007)*, Penang, Malaysia, 2007.
- [23] R. Kuhn, R. Kacker, Y. Lei, "Combinatorial Software Testing," *IEEE Transactions on Software Technologies*, August 2009, pp. 94-96.
- [24] D. M. Cohen, S. R. Dalal, A. Kajla, and G. C. Patton, "The Automatic Efficient Test Generator (AETG) System," in *Proc. of the 5th International Symposium on Software Reliability Engineering*, Monterey, CA, USA, 1994, pp. 303-309.
- [25] Y. Lei and K. C. Tai, "In-Parameter-Order: A Test Generation Strategy for Pairwise Testing," in *Proc. of the 3rd IEEE Intl. High-Assurance Systems Engineering Symp.*, Washington, DC, USA, 1998, pp. 254-261.
- [26] Y. Lei, R. Kacker, D. R. Kuhn, V. Okun, and J. Lawrence, "IPOG/IPOD: Efficient Test Generation for Multi-Way Software Testing," *Journal of Software Testing, Verification, and Reliability*, vol. 18, 2009, pp. 125-148.
- [27] J. Bach, "Allpairs Test Case Generation Tool," Available from: <http://tejasconsulting.com/open-testware/feature/allpairs.html>
- [28] "TConfig," Available from: <http://www.site.uottawa.ca/~awilliam/>.
- [29] "Jenny," Available from: <http://www.burtleburtle.net/bob/math/>.
- [30] "TVG," Available from: <http://sourceforge.net/projects/tvg>.
- [31] M. B. Cohen, "Designing Test Suites for Software Interaction Testing," PhD in Computer Science. New Zealand: University of Auckland, 2004.
- [32] M. Grindal, J. Offutt, and S. F. Andler, "Combination Testing Strategies: a Survey," *Software Testing Verification and Reliability*, vol. 15, 2005, pp. 167-200.
- [33] M. Grindal, B. Lindstrom, J. Offutt, S. F. Andler, "An Evaluation of Combination Strategies for Test Case Selection", Technical Report HS-IDA-TR-03-001, Department of Computer Science, University of Skövde, 2003.
- [34] D.R. Kuhn, R.N. Kacker and Y. Lei, *Combinatorial Coverage as an Aspect of Test Quality*, the *Journal of Defense Software Engineering*, 2014.
- [35] D.R. Kuhn, R.N. Kacker and Y. Lei, *Measuring and Specifying Combinatorial Coverage of Test Input Configurations*, *Innovations in Systems and Software Engineering: a NASA journal*, 2014.
- [36] J. Torres-Jimenez, I. Izquierdo-Marquez, *Survey of Covering Arrays*, 15th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC 2013), Timisoara, Romania, 23-26, 2013, pp. 20-27.
- [37] R.N. Kacker, D.R. Kuhn, Y. Lei, and J.F. Lawrence, *Combinatorial Testing for Software: an Adaptation of Design of Experiments*, *Measurement*, vol. 46, no. 9, 2013, pp. 3745-3752.
- [38] X. Niu, C. Nie, Y. Lei, A.T.S. Chan, *Identifying Failure-Inducing Combinations Using Tuple Relationships*, 2nd International Workshop on Combinatorial Testing (IWCT 2013), in *Proceedings of the Sixth IEEE International Conference on Software, Testing, Verification and Validation (ICST 2013)*, Luxembourg, March 18-22, 2013, pp. 271-280.
- [39] M.N. Borazjany, L.S.G. Ghandehari, Y. Lei, R.N. Kacker and D.R. Kuhn, *An Input Space Modeling Methodology for Combinatorial Testing*, 2nd International Workshop on Combinatorial Testing (IWCT 2013), in *Proceedings of the Sixth IEEE International Conference on Software, Testing, Verification and Validation (ICST 2013)*, Luxembourg, March 18-22, 2013, pp. 372-381.