ICIT 2015 The 7th International Conference on Information Technology
doi:10.15849/icit.2015.0007

ONLINE RECOGNITION SYSTEM FOR HANDWRITTEN ARABIC DIGITS

Mustafa Ali Abuzaraida
Computer Scince 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

doi:10.15849/icit.2015.0007

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:

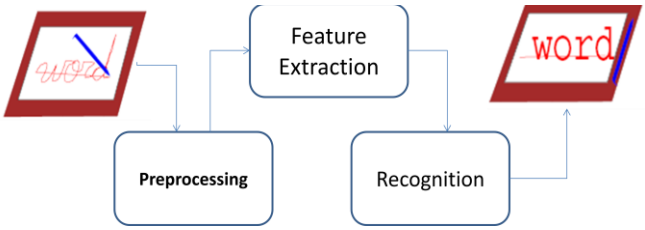

Figure 1. Typical phases of online Text Recognition System

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.

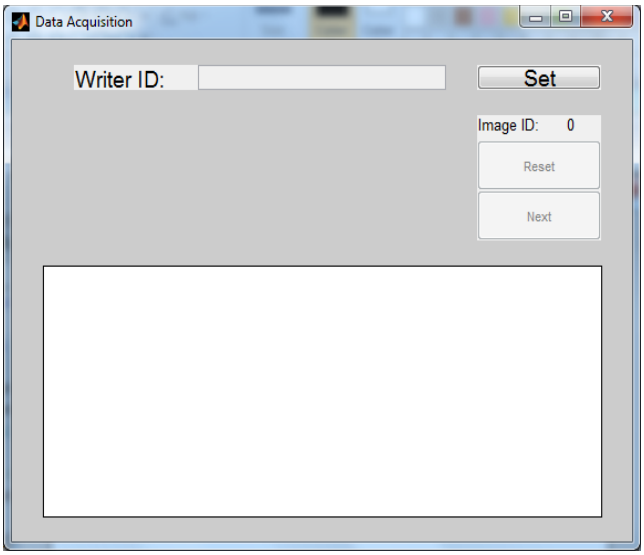

Figure 2. Data acquisition platform

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 (x_1, y_1) 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.

doi:10.15849/icit.2015.0007

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.

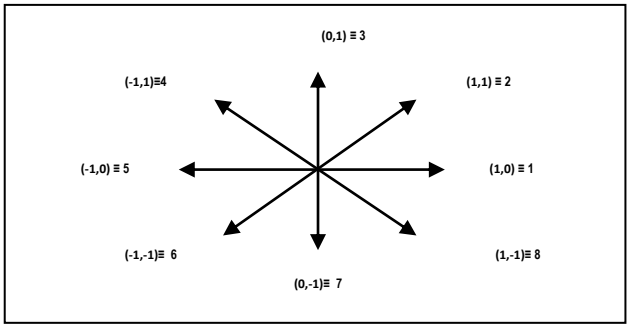

Figure 3. Freeman Chain Code

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.

doi:10.15849/icit.2015.0007

- 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 conceder 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

doi:10.15849/icit.2015.0007

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: IIUM 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.