ICIT 2015 The 7th International Conference on Information Technology
doi:10.15849/icit.2015.0103 © ICIT 2015 (http://icit.zuj.edu.jo/ICIT15)

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.

Gasparrini 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.

Ofli 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].

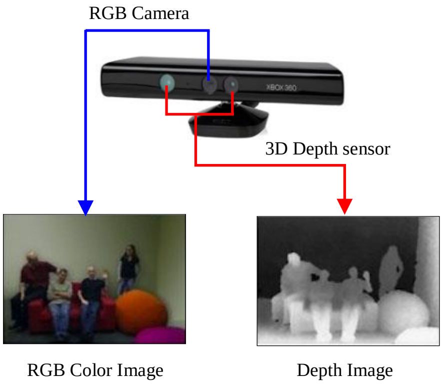


Fig. 1 RGB-D data captured by Kinect

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.

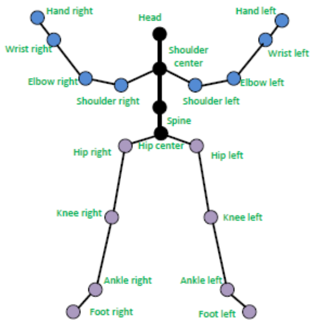


Fig. 2 [15] Skeleton joints detected by Microsoft SDK

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.

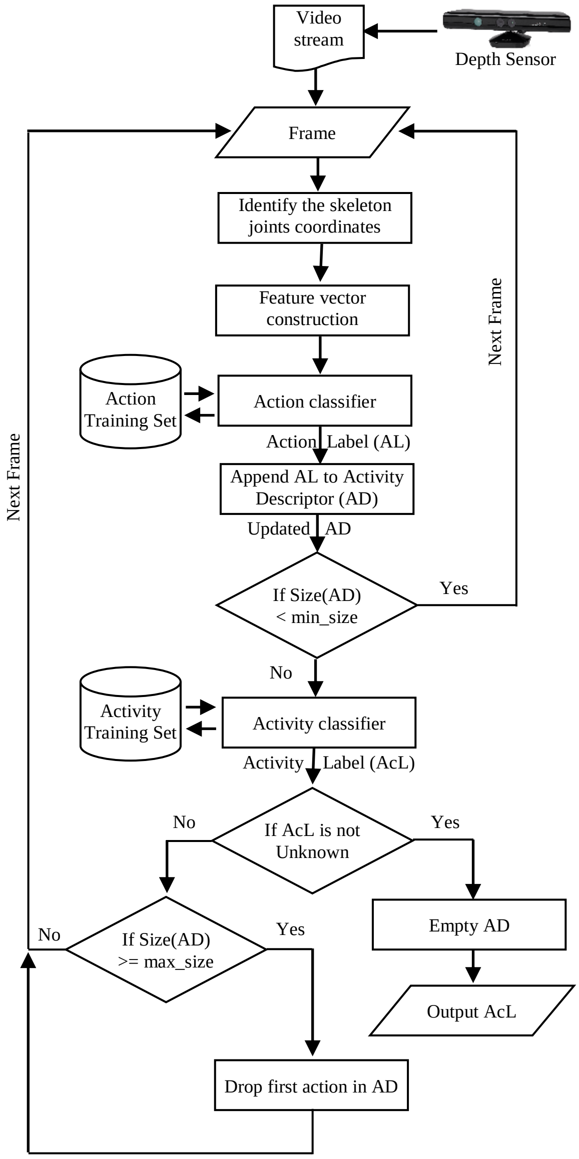


Fig. 3 The block diagram of the proposed system

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 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.

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.

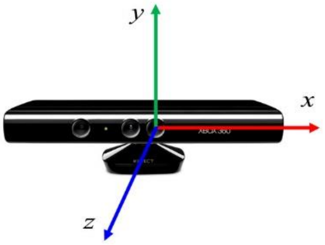
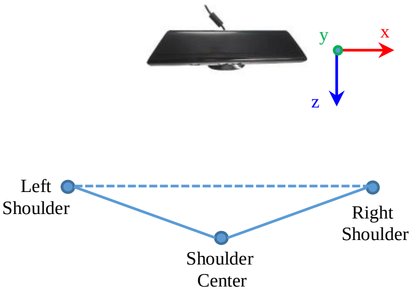
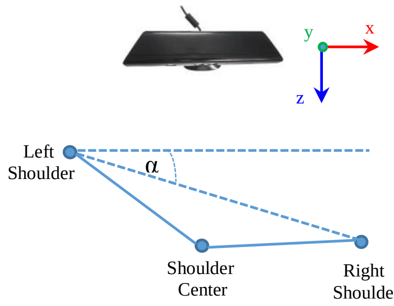


Fig. 4 Kinect Cartesian coordinate system



(a) The subject facing the depth camera

doi:10.15849/icit.2015.0103 © ICIT 2015 (http://icit.zuj.edu.jo/ICIT15)



(b) The subject at an angle with respect to 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):

$$\propto = \tan^{-1}\left(\frac{z_R - z_L}{x_R - x_L}\right) \qquad (1)$$

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):

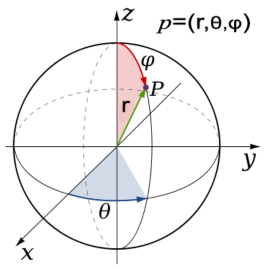
$$\begin{bmatrix} x_i' \\ y_i' \\ z_i' \\ 1 \end{bmatrix} = \begin{bmatrix} \cos\propto & 0 & \sin\propto & 0 \\ 0 & 1 & 0 & 0 \\ -\sin\propto & 0 & \cos\propto & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ z_i \\ 1 \end{bmatrix} \qquad (2)$$

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):

$$(x_i', y_i', z_i') = (x_i - x, y_i - y, z_i - z) \qquad (3)$$

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

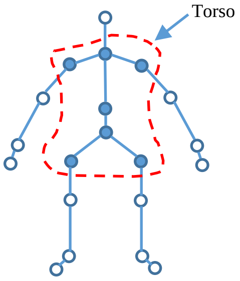


Fig. 7 Torso skeleton joints discarded from the feature vector

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 ∊ L, written as, C : s → l; l ∊ 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):

$$AcL = \arg\max_{i=1,2,\dots,N}\{P(AD|H_i)\} \qquad (4)$$

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/um/people/zliu/Action RecoRsrc/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/).

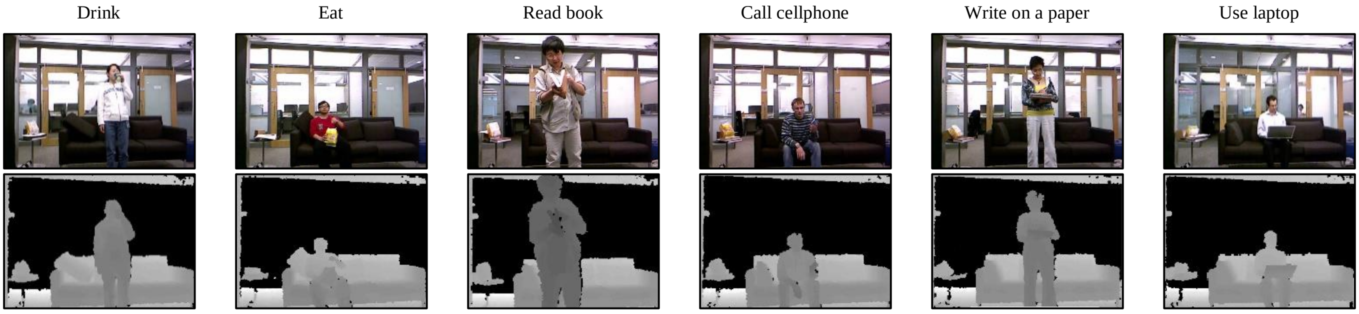


Fig. 8. Some example frames of MSR Daily Activity3D dataset, First row: RGB frames, Second row: depth images

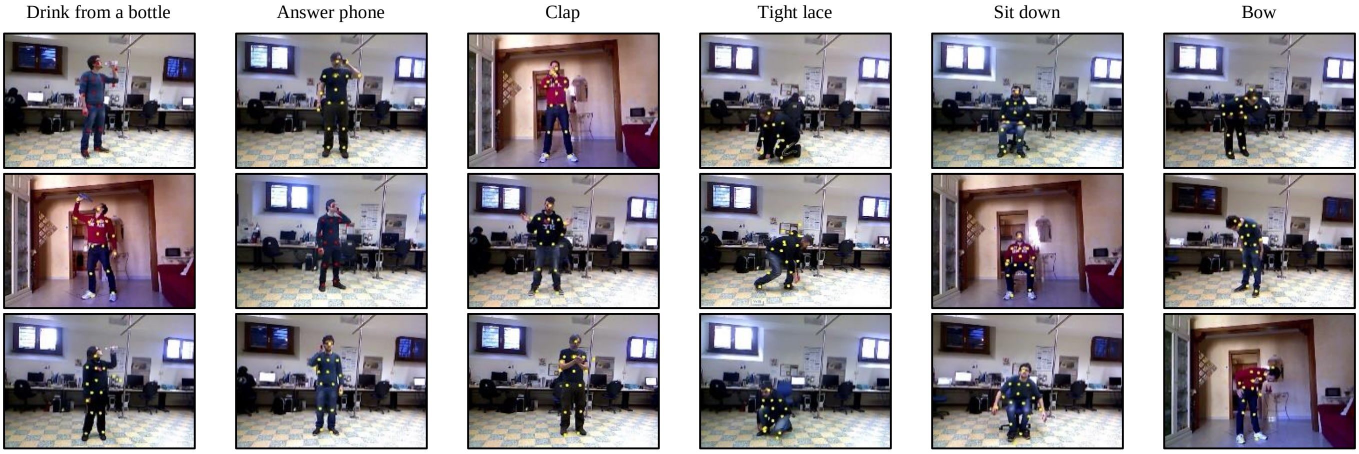


Fig. 9. Some example frames of Florence 3D dataset, each activity is performed by different subjects

drink	eat	Read Book	Call Cellphone	write	Use Laptop	Vaccum Cleaner	cheerUp	sitStill	tossPaper	Play Game	layDown	walk	Play Guitar	standUp	sitDown

doi:10.15849/icit.2015.0103 © ICIT 2015 (http://icit.zuj.edu.jo/ICIT15)

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			1	95	
bow		1	2			4	2	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 (IJCSET), 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] Maaike 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. Mahoora, 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 (IJAREEIE), 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 Chaaraouia, José Ramón Padilla-Lópeza, Pau Climent-Pérezb, 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

Weakly Aligned Multi-part Bag-of-Poses," In proceedings of the IEEE Computer Vision and Pattern Recognition Workshops (CVPRW), Portland, Oregon, USA, pp. 479-485, June 2013.