Depth Silhouettes Context: A New Robust Feature for Human Tracking and Activity Recognition based on Advanced Hidden Markov Model

Ahmad Jalal, Shaharyar Kamal, Daijin Kim
Kyung Hee University
Republic of Korea
ahmadjalal@postech.ac.kr

ABSTRACT: *In this paper, a depth camera-based novel approach for human activity recognition is presented using robust depth silhouettes context features and advanced Hidden Markov Models (HMMs). During HAR framework, at first, depth maps are processed to identify human silhouettes from noisy background by considering frame differentiation constraints of human body motion and compute depth silhouette area for each activity to track human movements in a scene. From the depth silhouettes context features, temporal frames information are computed for intensity differentiation measurements, depth history features are used to store gradient orientation change in overall activity sequence and motion difference features are extracted for regional motion identification. Then, these features are processed by Principal component analysis for dimension reduction and kmean clustering for code generation to make better activity representation. Finally, we proposed a new way to model, train and recognize different activities using advanced HMM. Each activity has been chosen with the highest likelihood value. Experimental results show superior recognition rate, resulting up to the mean recognition of 57.69% over the state of the art methods for fifteen daily routine activities using IM-Daily Depth Activity dataset. In addition, MSRAction3D dataset also showed some promising results.*

Keywords: Kinect camera, Depth silhouettes context features, Advanced Hidden Markov model, Human Activity Recognition (HAR)

Received: 14 July 2015, Revised 18 August 2015, Accepted 24 August 2015

1. Introduction

Recognizing human activities from videos have made greater attention of researchers in many research areas such as computer vision, human computer interaction, multimedia contents, image/signal processing and pattern recognition, which made access to wider applications like e-healthcare system, video surveillance, 3D games, security systems and smart homes systems [1]–[6].

Based on these domains, the factor which is most commonly used for person identification, verification and detection techniques is term as Human Activity Recognition (HAR) [7]–[11]. HAR is used to observe human movement, extract data in the form of signal or images and recognize indoor or outdoor activities from the sensor devices [12]. During HAR framework, there are three major issues. First is the type of sensor device, second is the data representation and last is the modeling, training or testing procedure. All of these issues underline the significance of human activity recognition in practical commercial areas [13].

Recently, many researchers in the field of human activity recognition based on a sequence of captured data from wearable sensors or video-based sensors. In wearable sensorbased HAR system, several devices (i.e., accelerometer and gyroscope) are attached with the human body during experimental scenarios to captured sequence of data. In [14], Junker et al. presented a method for spotting sporadically occurring gestures in a continuous data stream from bodyworn inertial sensors. They extract continuous sensor signals, uses a two-stage approach for the spotting task and exploiting the recognition capabilities via hidden Markov models. In [15], Liano et al. developed a method for human activity recognition that includes a hierarchical dynamic model, incorporating both inter-activity and intra-activity dynamics and exploiting the inherently dynamic nature of the problem to aid the classification task. This method uses raw acceleration and angular velocity signals which directly recorded by inertial sensors for feature extraction and recognized different activities using recognizer engine.

While, in case of video-based sensors, development in video cameras such as depth cameras (i.e., Kinect, bumblebee and Zeecam) have opened new opportunities of dealing in 3D data [16], [17]. In [18], Cheng et al. proposed a method to represent actions utilizing depth information and investigate the fusion of heterogeneous features from both color and depth sources to present a baseline performance for action recognition. In [19], Jalal et al. developed a life-logging HAR system that deals with motion features information such as magnitude and directional angular features of body joints information between consecutive frames to trained and recognized daily routine human activities. In [20], Sung et al. proposed an algorithm called hierarchical maximum entropy markov model (MEMM) which considers a person's activity as composed of a set of sub-activities. Then, two-layered graph structure is inferred by using a dynamic programming approach based on RGBD data from the Kinect sensor to recognize different activities. However, HAR system built by wearable sensors have certain limitations such as consumed high electric power, uncomfortable for the subject to wear for long time and relatively expensive in terms of energy consumption. On the other hand, video-based HAR system is quite feasible for real world application scenarios, therefore, our research work is focused on utilizing depth information for activity training and recognition.

In this paper, we propose a novel activity recognition framework based on depth silhouettes context features along with advanced HMMs using depth silhouettes. Based on depth maps, we segment human silhouettes from the noisy data using background subtraction technique. These depth silhouettes tracked human movements properly using frame differentiation scheme and extract features to locate spatiotemporal patterns from time-sequential activities. Also, these features deal with motion, translation distance and frame differentiation of human silhouettes which provide compact and sufficient information for human activity recognition. Then, based on the specific human silhouettes information (i.e., top, front, side and body parts motion) in different activities, the activities are divided into different classes. While, all features are mapped into codewords and recognize different human activities via advanced Hidden Markov model (HMM). We evaluate our method according to the standard experimental protocols definition on two different benchmark depth datasets: IM-DailyDepthActivity and MSRAction3D. Our results outperforms all published state of the art methods as shown in Table 1 and Table 2.

The organization of this paper is as follows. Section 2 introduces the system architecture of the proposed method from depth image acquisition, feature extraction by depth silhouettes contexts features and activity training and recognition by advanced HMMs. Section III explains experimental setting and results by considering proposed and state of the art methods. Finally, we provide a conclusion in Section 4.

2. Materials and Methods

Our proposed HAR system consists of the following steps: 1) depth map acquisition, 2) background subtraction, 3) human silhouettes identification, 4) feature extraction based on depth silhouettes context features, 5) clustering algorithm and vector quantization, and 6) activity training and recognition. Fig. 1 shows the overall flow of our proposed activity recognition system.

2.1 Depth Image Acquisition

To capture depth sequential maps, a depth camera (i.e., Kinect) is utilized and acquired a pair of RGB and depth maps [21]. To extract the noisy background surfaces from the depth map [8], we applied foreground, background and depth differentiation method as

$$f_t(x,y,z) = \begin{cases} 1 \text{ if } b_t(x,y,z) - d_t(x,y,z) > T_{value} \\ 0 \qquad otherwise \end{cases} \tag{1}$$

where $f_t(x,y,z)$, $b_t(x,y,z)$ and $d_t(x,y,z)$ are the foreground, background and depth intensity pixel values at time t and T_{value} is a positive threshold value. Also, to detect the human moving silhouettes, temporal depth intensity distribution is applied to obtain the depth human silhouettes from the noisy regions.

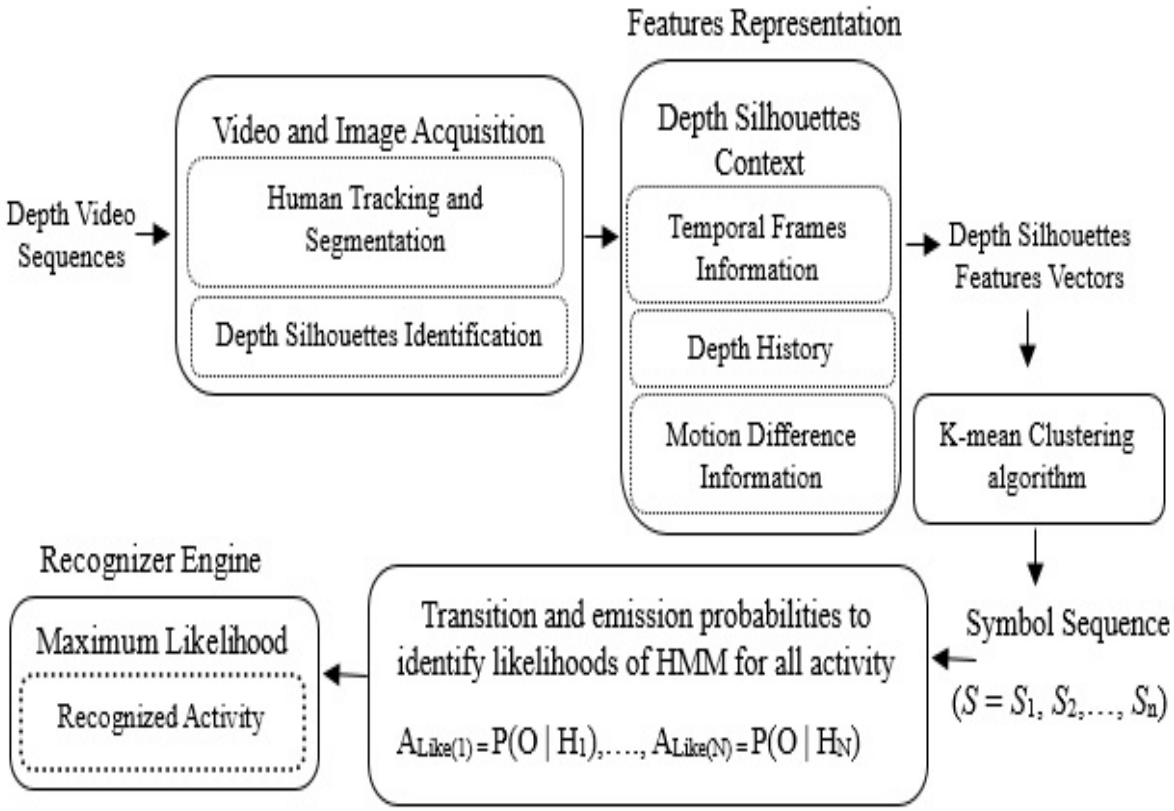

Figure 1. Overall flow of the proposed human activity recognition system

Figure 2. Temporal frames information features are applied using 2D frontal frames projections

Figure 3. Samples images applied by depth history features from different RGB-D video datasets

2.2 Feature Representation and Extraction Using Depth Silhouettes Context features

In this section, we extract features from full-body silhouettes based on depth images. However, a set of feature representation techniques such as temporal frames information, depth history and motion difference information are derived to augment together having spatial and temporal depth silhouettes characteristics.

1) Temporal frames information: From a sequence of depth images, we extract motion intensity difference of human silhouettes by considering the difference d_f among two consecutive frames t and t -1 as

$$d_f = |f^i_t - f^i_{t-1}| \tag{2}$$

To consider additional temporal information from depth silhouettes, each depth frame d_f is projected into 2D frontal planes using orthogonal cartesian planes as shown in Figure. 2. However, the information from the frontal view is quite effective and prominent which make it unique in order to increase the feature accuracy during classification and recognition.

2) Depth history features: To emphasize on spatiotemporal region in overall sequence, we considered the shape information along with motion region of each body parts in an activity. Therefore, we calculated the gradient orientation, pixel intensity in-between initial frame till final frame and Mahalanobis distance for matching the input activity from the stored templates [22]. Figure. 3 explains the detail description of depth history features using depth dataset.

Figure 4. Motion difference information features are extracted from human depth silhouettes

3) Motion difference information: To observe the local motion features of specific body parts, we calculate the distance of upper, middle and lower portions *sp* of human silhouettes by considering Euclidean distance and cosine distance of human silhouettes *s* and *r* [23], [24]. Thus, to identify the larger α and smaller β spatial distance motion in-between consecutive human silhouettes are expressed as

$$d(I) = \begin{cases} 1 - min(sp_s(I), sp_r(I)) & if\ d_i(s,r) < \beta \\ \alpha & if\ d_i(s,r) \geq \beta \end{cases} \tag{3}$$

$$D(s,r) = \sum_{I=1}^{n} d(I) \tag{4}$$

Therefore, depth silhouettes context features are capable of providing optimal clustering for all different activities and separate them based on local motion [2], spatial/temporal characteristics and matching ability.

2.3 Features Dimension Reduction and Symbol Representation

These full-body spatial/temporal features space consists of larger number of features dimension, thus, Principal component analysis (PCA) is defined here to extract global information from all activities data and approximate the higher features dimension data into lower dimensional features [10], [16]. In this work, 500 principal components (PCs) are used to process the activity data and are expressed as $F_{PC} = \widetilde{M}_i E_{top}$ where F_{PC} is the PCA projection of feature vectors, $\widetilde{M}_i$ is the zero mean vector and E_{top} is the top eigenvectors indicating higher variance [25]. Figure. 5 shows the top 500 eigenvalues with respect to their eigenvectors.

However, the size of depth silhouettes context features are quite higher dimension, therefore, we optimize each feature vector upto 1×500.

Then, they are symbolized by the codebook that is generated from k-mean clustering algorithm. All features are represented by the codes that minimize the sum of squared distance between the input depth silhouettes features and prototype vectors. It mainly based on clustering quality and iteration number with respect to initial *k* cluster centers. In each cluster, *N* samples that are nearest to their center points acting as specific same class (i.e., activity label). However, these activity feature codes are generated per each sequence with an overlap limit and maintained by buffer strategy [26], [27]. Figure. 6 shows the internal concept of dimensional structure and code selection of proposed features.

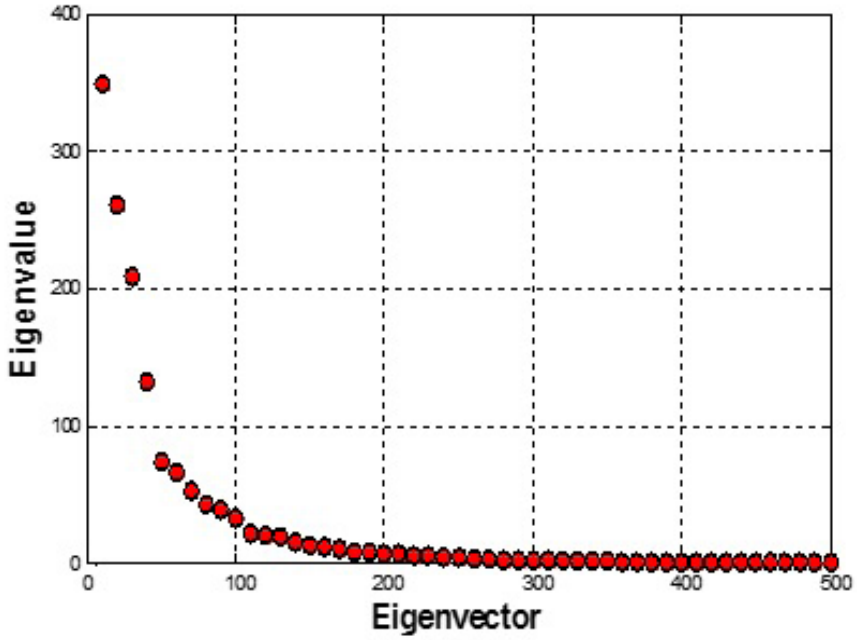

Figure 5. Top 500 eigenvalues are identified with respect to its eigenvectors

2.4 Training and Recognition via Advanced Hidden Markov Model (A-HMM)

For training and recognition, we introduced advanced HMM method. However, in conventional HMM approach, HMM consists of finite states having transition probability and symbol observation probability. It includes the overall human silhouettes information which contain redundant information such as static body regions (i.e.,torso, chest and hips) and unmovable body parts (i.e., arms, thigh and forearm). Such kind of unnecessary information causes reduction at overall performance of accuracy results. Therefore, advanced HMMis introduced which focused on specific and active areas of human body parts such as arms, legs, head, feet and shoulders. Thus, active feature regions (i.e., front, side, top and motion body parts areas) are augmented

together to build a single HMM having unique information for each activity. Figure. 7 shows active features regions of overall human silhouettes to calculate specific likelihood of each activity.

$$L_{activity} = arg\ max_{k} \{ P (O | \lambda_k) \quad (5)$$

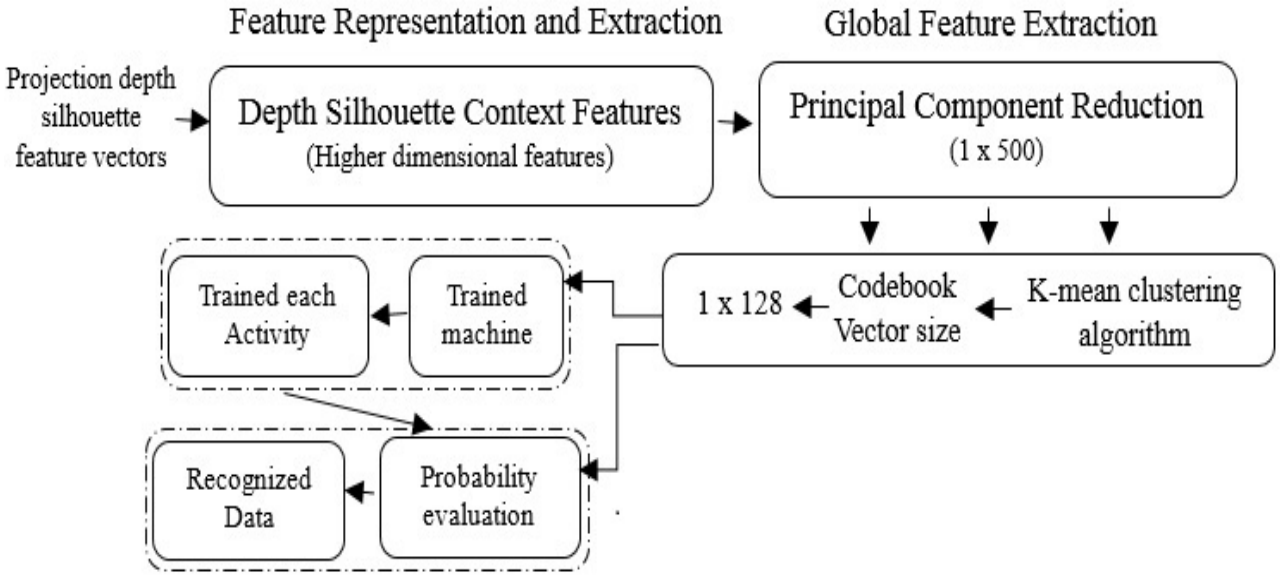

Figure 6. Overall procedure of feature dimensional projection over depth silhouettes

Finally, maximum likelihood value is used to choose the desired activity during testing.

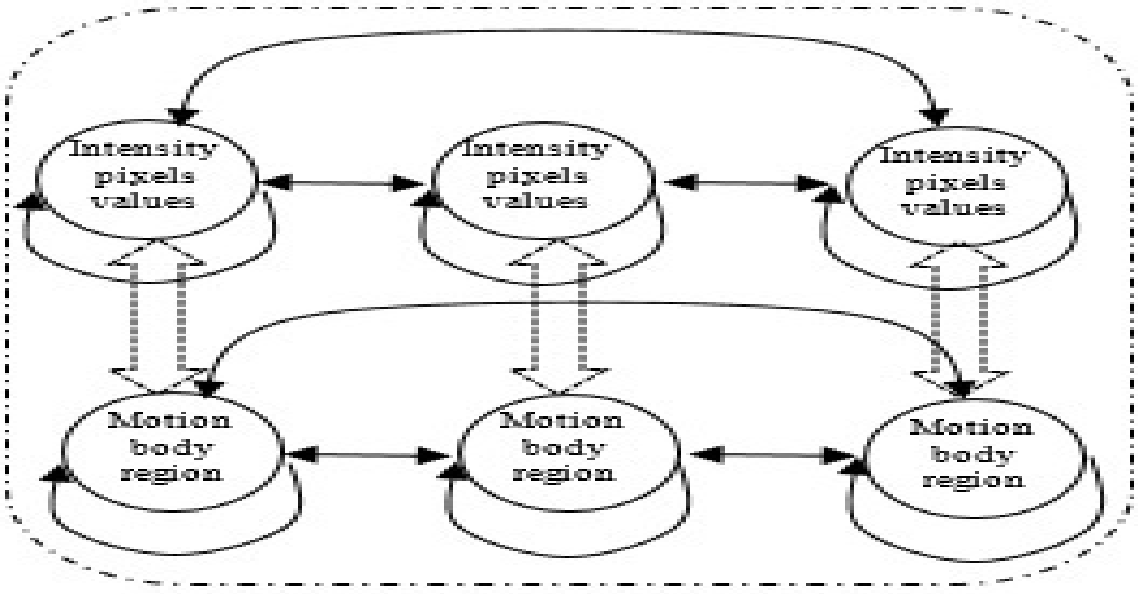

Figure 7. Transition and emission probabilities distribution based on advanced HMM for all active human silhouettes areas

3. Experimental Results

In this section, we explain the experimental settings of our own annotated dataset (i.e., IM-DailyDepthActivity) and public depth dataset (i.e., MSRAction3D). A comparison of results of recognition accuracy is also considered between state of the art methods and the proposed method.

3.1 Annotated dataset as IM-DailyDepthActivity

According to our knowledge, there is no benchmark dataset available for online activity recognition using depth camera. For that reason, we collected a new online continuous activity dataset called IM-DailyDepthActivity dataset [28] which includes daily life activities. It contains fifteen activity categories: sit down, both hands waving, phone conversation, kicking, reading an article, throwing, bending, clapping, right hand waving, take an object, exercise, eating, boxing, cleaning and stand up, respectively. All activities are captured in university environments (i.e., labs and classrooms). Figure. 8 gives an example of all fifteen activity categories. However, the dataset includes 45 segmented videos of each activity for training and 30 unsegmented continuous videos for testing performed by 15 different subjects at different time interval. All sample files include RGB frames, depth maps, silhouettes labeling and skeleton joints information.

We compare our depth silhouettes context features approach with the approach using state of the art features as R transform [29], [30] where r transform features computed a 1D profile of a depth silhouettes along specified view direction for all daily human activities. These projections established a mapping between the domain produced by depth coordinates and the respective angles. While, eigenjoints method [31] deals with a new type of features based on position differences of joints and EigenJoints, which combine action information including static posture, motion, and offset. Then, Naive-Bayes-Nearest-Neighbor (NBNN) classifier is used as recognizer engine.

Methods	Recognition Accuracy
R transform features	[30] 33.74
EigenJoints	[31] 40.35
Depth Silhouettes Context features	57.69

Table 1. Recognition accuracy comparison using IMDailyDepthActivity dataset

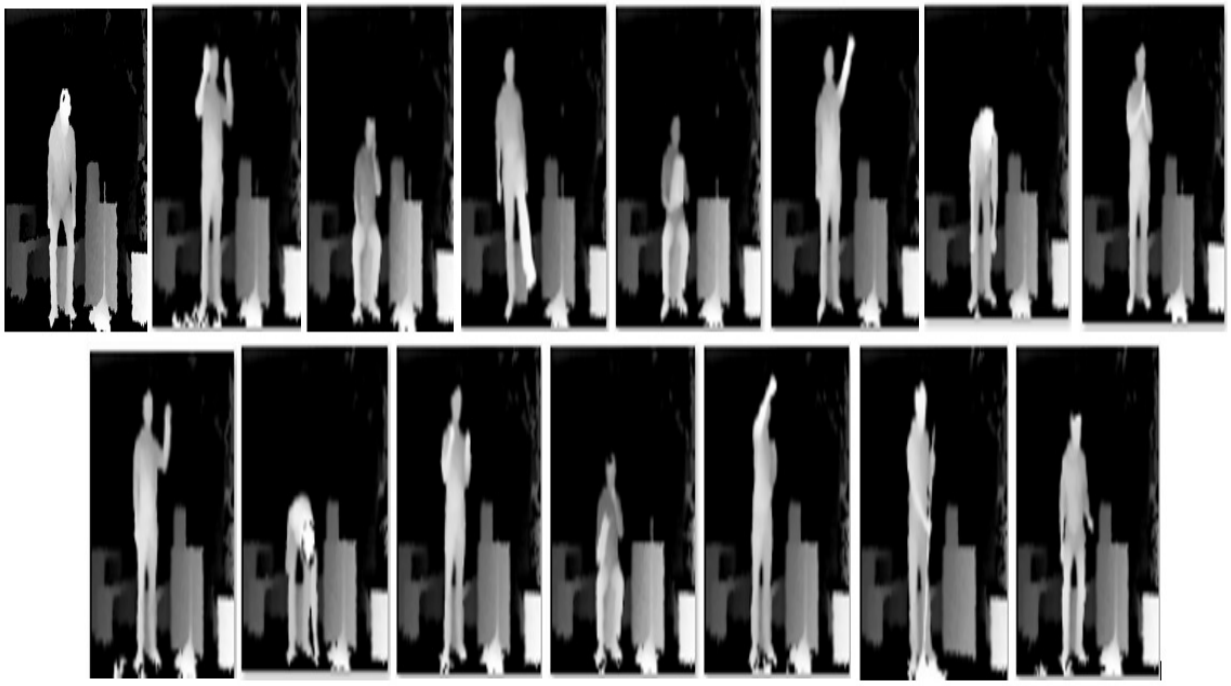

Figure 8. Sample depth silhouettes images of annotated IM-DailyDepthActivity dataset

Table 1 shows the recognition results of state of the art methods (i.e., R transform features and EigenJoints) and proposed body joints features using depth activity silhouettes respectively whereas the proposed method shows significantly superior mean recognition of 57.69% over that of the state of the art methods (i.e., 33.74% and 40.35%).

Figure 9. Sample depth images of MSRAction3D dataset

3.2 Public Dataset as MSRAction3D

The MSRAction3D dataset was captured with a depth sensor (i.e.Kinect device) by the Microsoft Researcher team. It includes 20 different action types as: high arm wave, horizontal arm wave, hammer, hand catch, draw x , forward punch, jogging, two hand wave, high throw, draw tick , draw circle, hand clap, bend, side boxing, forward kick , side kick , tennis swing, tennis serve, golf swing and pickup & throw. The dataset consists of 567 depth map sequences performed by 10 subjects. Also, the background of this dataset is clean and the human silhouettes are available in each frame. This dataset is quite challenging due to similar postures of different action especially hands and legs movements. Several samples of MSRAction3D dataset are shown in Figure. 9.

To compare the proposed features method at MSRAction3D, we considered the solutions reported as the state of the art methods [32]–[34]. In [32], Martens and Sutskever demonstrated the Hessian-Free optimizer augmented with the structural damping approach, is capable of robustly training Recurrent neural networks (RNNs) to solve tasks that exhibit long term dependencies during action sequences. While, dynamic temporal warping [33] proposed new methods for automatic classification and retrieval of motion capture data facilitating the identification of logically related motions scattered in action dataset. In [34], Xia et al. presents a novel approach for human action recognition with histograms of 3D joint locations (HOJ3D) as a compact representation of postures. These HOJ3D are projected over LDA and recognized using HMM. It is clearly shown that the proposed method achieved improved recognition rate as 83.92% as compare to the state of the art methods as 42.5%, 54.0% and 79.0% respectively.

Methods	**Recognition Accuracy**
Recurrent Neural Network	[32] 42.5
Dynamic Temporal Warping	[33] 54.0
HOJ3D	[34] 79.0
Depth Silhouettes Context features	83.92

Table 2. Recognition accuracy comparison using MSRAction3D dataset

4. Conclusions

In this paper, a human activity recognition system has been proposed utilizing depth silhouettes context features along with advanced HMMs using depth sensor. Based on proposed features obtained from the time-sequential activity video frames, we configure spatio-temporal properties of features, specific body parts motions, gradient orientation and local motion of human silhouettes. These features are mapped into codewords to generate a sequence of discrete symbols for training advanced HMMs based on specific information. For recognition, one activity has been chosen with the highest likelihood among other

activities. Our proposed feature method provides higher recognition accuracy performance over the state of the art methods using both depth images datasets. Our marker-free activity recognition system should be practical in different consumer applications such as pedestrian detection, 3D video games, smart homes and patient monitoring systems.

In future, we have planned to merge the proposed features with the joint points information in order to improve the effectiveness of our system. Also, we will use these features over other datasets like person-to-person interaction or human object interaction.

References

[1] Chaminda, H., Klyuev, V., Naruse, K. (2012). A smart reminder system for complex human activities, *In*: International Conference on Advanced Communication Technology (ICACT) p. 235–240.

[2] Jalal, A., Kim,S. (2006). Global security using human face understanding under vision ubiquitous architecture system. World Academy of Science, Engineering, and Technology, 13.

[3] Holte, M., Cuong, T., Trivedi,M., Moeslund,T. (2012). Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments, *Selected Topics in Signal Processing, IEEE Journal of*, 6 (5) 538–552.

[4] Jalal, A., Shahzad, A. (2007). Multiple facial feature detection using vertex-modeling structure, *In*: IEEE Computer Society Conference on Interactive computer aided learning, p. 1–7. IEEE.

[5] Jalal, A., Zeb, M. A. (2008). Security enhancement for e-learning portal, *International Journal of Computer Science and Network Security,* 8 (3) 41–45.

[6] Niu, W., Long, J.,Han, D. ,Wang, Y. (2004).Human activity detection and recognition for video surveillance, in IEEE International Conference on Multimedia and Expo, p. 719–722. IEEE.

[7] Veeraraghavan, A., Roy-Chowdhury, A. K., Chellappa, R. (2005). Matching shape sequences in video with applications in human movement analysis, *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* 27 (12) 1896–1909.

[8] Jalal,A., Kim, J. T., Kim, T.-S. (2012). Human activity recognition using the labeled depth body parts information of depth silhouettes, *In*: Proceedings of the 6th international symposium on Sustainable Healthy Buildings, p. 1–8.

[9]Arie, J. B., Wang, Z., Pandit, P., Rajaram, S. (2002).Human activity recognition using multidimensional indexing, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 24 (8) 1091– 1104.

[10] Jalal, A., Sarif, N.,Kim, J. T., Kim, T.-S. (2013).Human activity recognition via recognized body parts of human depth silhouettes for residents monitoring services at smart home, Indoor and Built Environment, 22 (1) 271–279.

[11] Turaga, P., Chellappa, R., Subrahmanian, V. S. Udrea, O. (2008). Machine recognition of human activities: A survey, *Circuits and Systems for Video Technology, IEEE Transactions on,* 18 (11) 1473– 1488.

[12] Tolstikov, A., Phua, C., Biswas, J., Huang, W. (2011). Multiple people activity recognition using mht over dbn, *In*: Toward Useful Services for Elderly and People with Disabilities, p. 313–318. Springer,.

[13] Jalal, A.. Kim, Y., Kim, D. (2014). Ridge body parts features for human pose estimation and recognition from rgb-d video data, *In* : Proceedings of the IEEE Conference on computing, communication and networking technologies, p. 1–6. IEEE.

[14] Junker, H., Amft, O., Lukowicz, P., Troster, G. .(2008). Gesture spotting with body-worn inertial sensors to detect user activities, *Pattern Recognition*, 41 (6) 2010–2024.

[15] Liano, B., Mahony, N., Rodriguez, A. (2012). Hierarchical dynamic model for human daily activity recognition, *In*: Proceedings of the International Conference on Bio-inspired systems and signal processing, p. 61–68, 2012.

[16] Jalal, A., Kim, J. T., Kim, T.-S. (2012).Development of a life logging system via depth imaging-based human activity recognition for smart homes, in Sustainable Healthy Buildings, International Symposium on, p. 91–95.

[17] Jalal, A., Uddin, M., Kim, J. T., Kim, T.-S. (2011). Daily human activity recognition using depth silhouettes and r transformation for smart home, *In*: Proceedings Smart Homes Health Telematics, p. 25–32. Springer.

[18] Cheng, Z., Qin, L., Ye, Y., Huang, Q., Tian, Q. (2012). Human daily action analysis with multi-view and color-depth data, *In*: Proceedings of the International Conference on Computer Vision, p. 52–61.

[19] Jalal, A., Kamal, S., Kim, D. (2014). A depth video sensor-based lifelogging human activity recognition system for elderly care in smart indoor environments, *Sensors*, 14 (7) 11 735–11 759.

[20] Sung, J. , Ponce, C., Selman, B., Saxena, A. (2012). Unstructured human activity detection from rgbd images, *In:* Robotics and Automation (ICRA), 2012 *IEEE International Conference on*, p. 842–849. IEEE.

[21] Jalal, A., Kamal, S. (2014). Real-time life logging via a depth silhouettebased human activity recognition system for smart home services, *In*: Proceedings of the IEEE International Conference on Advanced Video and Signal-based Surveillance, p. 74–80. IEEE.

[22] Jalal, A.,Kim, Y. (2014). Dense depth maps-based human pose tracking and recognition in dynamic scenes using ridge data, in Proceedings of the IEEE International Conference on Advanced Video and Signal-based Surveillance, p. 119–124. IEEE.

[23] Jalal , A., Uddin, I. (2007), Security architecture for third generation (3g) using gmhs cellular network, *In*: Emerging Technologies, 2007. ICET 2007. International Conference on, p. 74–79. IEEE.

[24] Jalal , A., Zeb, M. A. (2007).Security and qos optimization for distributed real time environment, *In:* Computer and Information Technology, 2007. CIT 2007. 7th IEEE International Conference on, p. 369–374. IEEE.

[25] Jalal, A., Kamal, S., Kim,D. (2015). Shape and motion features approach for activity tracking and recognition from kinect video camera, *In* : Proceedings of the International Conference on Advanced Information Networking and Applications Workshops, p. 445–450.

[26] Jalal, A., Rasheed, Y. A. (2007). Collaboration achievement along with performance maintenance in video streaming, in Interactive Computer Aided Learning, IEEE Conference on,

[27] Jalal, A., Kim, S. (2005). The mechanism of edge detection using the block matching criteria for the motion estimation, in Human Computer Interaction, 2005. HCI 2005. *In*: Proceedings of Conference on, 484– 489.

[28] Jalal, A.(2015). IM-DailyDepthActivity dataset, imlab.postech.ac.kr/ databases.htm. [Online; accessed 14-April-2015].

[29] Jalal, A., Uddin, M. Z., Kim, J. T., Kim, T.-S. (2012). Recognition of human home activities via depth silhouettes and r transformation for smart homes, *Indoor and Built Environment*, 21 (1) 184–190.

[30] Jalal, A., Uddin,M. Z. , Kim, T.-S. (2012).Depth video-based human activity recognition system using translation and scaling invariant features for life logging at smart home,*Consumer Electronics, IEEE Transactions on*, 58(3) 863–871.

[31] Yang , X., Tian, Y. (2012).Eigenjoints-based action recognition using naive-bayes-nearest-neighbor, *In*: Proceedings of the IEEE International Conference on Computer vision and pattern recognition workshops, p. 14–19.

[32] Martens, J., Sutskever, I. (2011). Learning recurrent neural networks with hessian-free optimization, *In:* International Conference on Machine learning, p. 1033–1040.

[33] Muller, M., Roder, T. (2006). Motion templates for automatic classification and retrieval of motion capture data, in ACM SIGGRAPH/Eurographics symposium on computer animation, p. 137–146.

[34] Xia, L., Chen,, C., Aggarwal, J. (2012). View invariant human action recognition using histograms of 3d joints, *In*: Proceedings of the IEEE International Conference on Computer vision and pattern recognition workshops, p. 20–27.