A Robust Approach for Human Activity Recognition Using 3-D Body Joint Motion Features with Deep Belief Network

Md. Zia Uddin[1] and Jaehyoun Kim[2]
[1]Department of Informatics, University of Oslo
Oslo, Norway
[2]Department of Computer Education, Sungkyunkwan University
Seoul, South Korea
[e-mail: mdzu@ifi.uio.no, jaekim@skku.edu]
*Corresponding author: Jaehyoun Kim

Received August 29, 2016; revised December 25, 2016; accepted February 9, 2017;
published February 28, 2017

Abstract

Computer vision-based human activity recognition (HAR) has become very famous these days due to its applications in various fields such as smart home healthcare for elderly people. A video-based activity recognition system basically has many goals such as to react based on people's behavior that allows the systems to proactively assist them with their tasks. A novel approach is proposed in this work for depth video based human activity recognition using joint-based motion features of depth body shapes and Deep Belief Network (DBN). From depth video, different body parts of human activities are segmented first by means of a trained random forest. The motion features representing the magnitude and direction of each joint in next frame are extracted. Finally, the features are applied for training a DBN to be used for recognition later. The proposed HAR approach showed superior performance over conventional approaches on private and public datasets, indicating a prominent approach for practical applications in smartly controlled environments.

Keywords: Depth video, 3-D body joints, Deep Belief Network (DBN).

A preliminary version of this paper was presented at APIC-IST 2016, and was selected as an outstanding paper.

https://doi.org/10.3837/tiis.2017.02.028 **ISSN : 1976-7277**

1. Introduction

In recent years, video-based Human Activity Recognition (HAR) has been getting a lot of attentions by many researchers in image processing, computer vision, pattern recognition, and human computer interaction (HCI) [1]. A key reason is to utilize HAR system in practical applications such as smart surveillance and healthcare systems. Basically, an HAR system consists of extracting features and applying them to compare with the feature database to see the similarities with features of different clusters already stored in the database. Thus, feature extraction, activity modeling, and recognition are the crucial parts of an HAR system. There are various inputs can be applied in to a robust HAR system. Among whhich, video sensors are very much used for activity recognition from images and hence it becomes a challenging task as video-based HAR considers whole body movement of human being in images but not only rigid regions such as hands in hand gesture recognition. Though there are many computer vision researchers who have been working on video-based HAR systems due to their prominent applications but accurate recognition of human activities in this regard is still considered to be a very big concern for most of them.

2. Related HAR Studies

For human activity feature extraction from images, 2-D binary shapes seem to be very common for feature extraction in HAR [1]-[3]. In [2], the authors statrted dsicussing abouot global shape feature representation such as Principal Component (PC) features of binary shapes to represent several activities. As PC features represent the global features and result in poor activity recognition performance, local body shape features such as Independent Component (IC) features were adopted later for better HAR. Later on, they showed the superiority of IC-based binary body shape features for HAR over the PC-based ones [2]. Though binary shapes are easy to implement for HAR, they have some limitations. For instance, binary shapes cannot represent difference between the near and distant body parts. However, depth information of body shapes can handle this problem and one could utilize depth body shape-based for robust HAR such as in [3]. Though depth shapes seem to be better than binary ones but different body parts cannot be separated if one consider the whole body shape and hence indicating segregation of different body parts to get the joints in the image as the human body is of different parts connected together. One can get stronger features from body joints than whole body features that may represent more robust HAR. Depth information-based pattern recognition has attracted a lot of researchers in the pattern recognition and computer vision fields for various applications such as human motion analysis [4]-[9]. Along with HAR, body part segmentation is also grabbing good attentions by computer vision researchers [10]-[14]. In [10], the authors used k-means algorithms for body part segmentation. In [11], the authors considered upper body part segmentation for estimating human pose. In [12], the authors adopted a manual body part segmentation approach to get body joints to be applied in recognition of gaits.

For training and recognition of time-sequential features, Hidden Markov Model (HMM) has been considered as a basic tool in decoding time-quential image analysis such as [15], [16]. For pattaern recognition from images, depth images is atracting a lot of researchers for varous practival applications these days [17]–[19]. In [17], the authors applied depth-image for human-activity recognition. In [18], the authors showed a unique depth image-based activity analysis based on surface-orientation histograms. In [19], the authors used motion energies on depth images for activity analysis. In [20], the authors analyzed object segmentation from RGB and depth images for activity analysis. In [21], the authors used Maximum Entropy Markov Model (MEMM) for human acitvity recognition where they used two layers of models. In [22], the authors used analyzed two interacting hands in depth images for human activity represenation. Some researchers also focused on visual gestural languages such as American Sign Language (ASL) [23]. For instance, in [23], the authors showed textual representations of continuous visual sign languages from depth information. In [24], the authors analyzed features from body joints from noisy depth images where stereo cameras were deployed for obtaining depth images and then, the features were applied with hidden Markov models (HMMs) for activity recognition. These days, Deep Neural Network (DNN) has gained much attentions by machine leraning researchers since it can generate some features from raw inputs [25]. Hilton et al. proposed Deep Belief Network (DBN), an improved version of DNN utilizing Restricted Boltzmann Machine (RBM) [26].

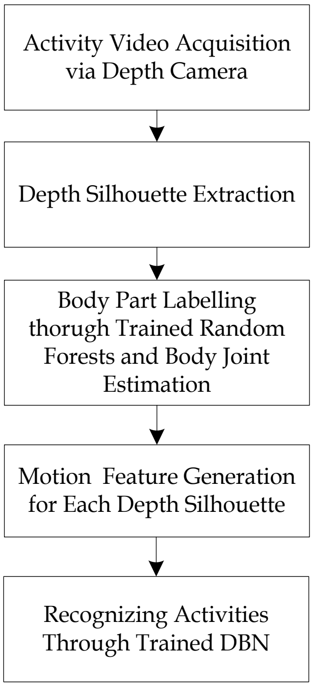

Activity Video Acquisition via Depth Camera
Depth Silhouette Extraction
Body Part Labelling thorugh Trained Random Forests and Body Joint Estimation
Motion Feature Generation for Each Depth Silhouette
Recognizing Activities Through Trained DBN

Fig. 1. Architecture of the proposed HAR system.

In this work, an efficient approach is described for HAR using body joints' motion features with DBN. For training activity, after extracting the body joints, motion features are generated from each depth image of the activity videos. Then, the feature sequences from the activity videos are augmented to train DBN. For testing an activity in a depth video, the augmented motion features from that video are applied on the trained DBN.

3. Proposed HAR Methodology

The proposed approach consists of video acquisition, segmentation of body parts through random forest, feature generation, and modeling DBN as depicted in **Fig. 1**. Kinect, a commercial camera is used to obtain the depth images of activities [27] and the body shape is extracted from every depth image..

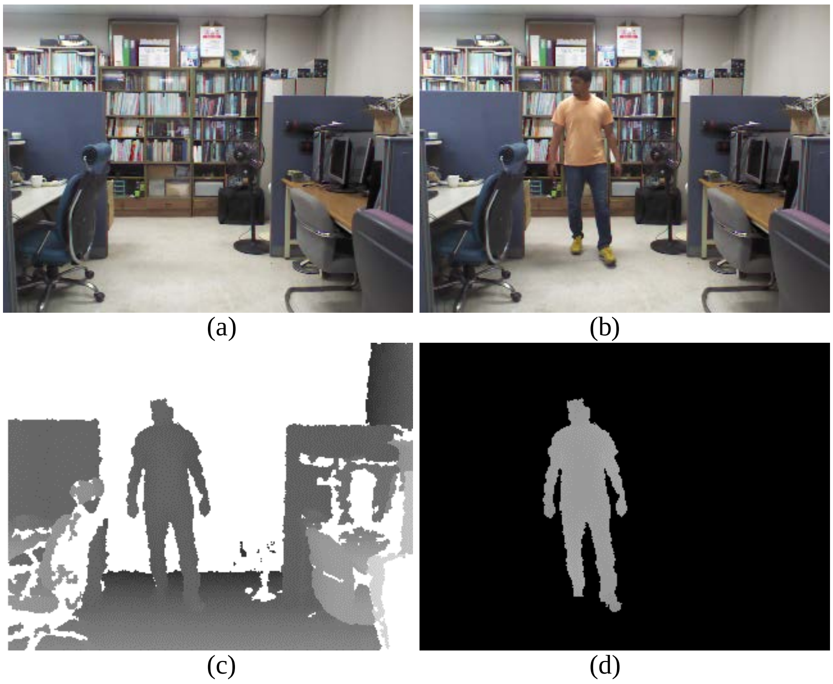

(a) (b) (c) (d)

Fig. 2. (a) A background scene, (b) a scene with a human, (c) corresponding depth map of (b), and (d) depth map of the subject extracted from (c).

3.1 Silhouette Extraction

Most of the background pixels in an image on our daily applications contain very high distance values and hence, considering thresholds on the depth values obtained through distance information in the image can make the human body silhouette extraction easy. In this regard, Gaussian Mixture Model (GMM) can be a good candidate to update the background for background subtraction to extract depth body silhouettes [28]. In this work, the background pixel probability is represented as

$$P(x)=\sum_{i=1}^{K} B_i * \eta(x,\mu,V) \tag{1}$$

where B_i is the weight of the i^{th} Gaussian mixture, μ is the mean, and V variance for that mixture. The Gaussian probability density function G can be modelled as:

$$G(x,\mu,V)=\frac{1}{(2\pi V)^{\frac{1}{2}}} e^{-\frac{1}{2V}(x-\mu)^2} \tag{2}$$

So, the background is updated based on the previous pixels over the time and once the background subtraction is done, human body is tracked in the image using shape context descriptors. To continue the matching for tracking a known body shape, the shape context cost C_n for y^{th} unknown region is obtained by comparing i^{th} point cost $L_{i,r}$ in the known silhouette with the unknown region's random points in the next frame as

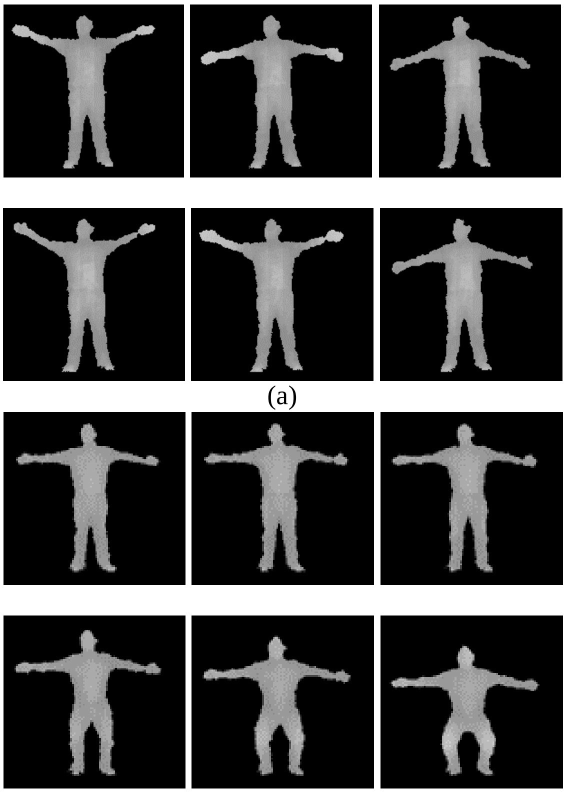
$$L_{i,r} = Min\left(\frac{1}{2}\sum_{l=1}^{u}\underset{j=1}{\overset{n}{\frac{(h_i(k)-h_j(k))^2}{h_i(k)+h_j(k)}}}\right) \tag{3}$$

$$C_n = \sum_{i=1}^{k} L_i \tag{4}$$

where h_i and h_j are log-polar normalized histograms of i^{th} point from k points in the known silhouette and j^{th} point from n points in y^{th} unknown region in next frame. All the costs of the polar bins are then added to get the total cost for matching a known silhouette and unknown region pair. Finally, the pair that returns the least cost is chosen the matching to select. The decision of matching can be taken as

$$Matched\_Shape\_Decision = \arg\min_{r=1}^{Y}(C) \tag{5}$$

where Y represents the number of regions to be matched. On the other contrary, matching cost of a body silhouette in a frame with wrong regions in next frame must return high cost and hence the matching is ignored. **Fig. 2** shows a sample of background scene, a scene with a human, the depth scene, and at last, the extracted depth map of the subject in the scene respectively. **Fig. 3** represents a sequence of depth body shapes from both hand waving and sitting down activities respectively


(a)

(b)

Fig. 3. Sequence of depth body silhouettes from (a) both hand waving and (b) sitting down activities.

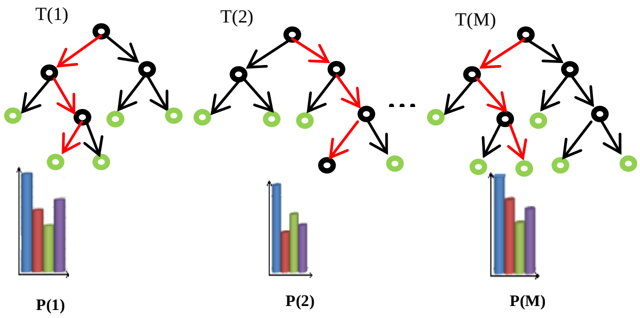
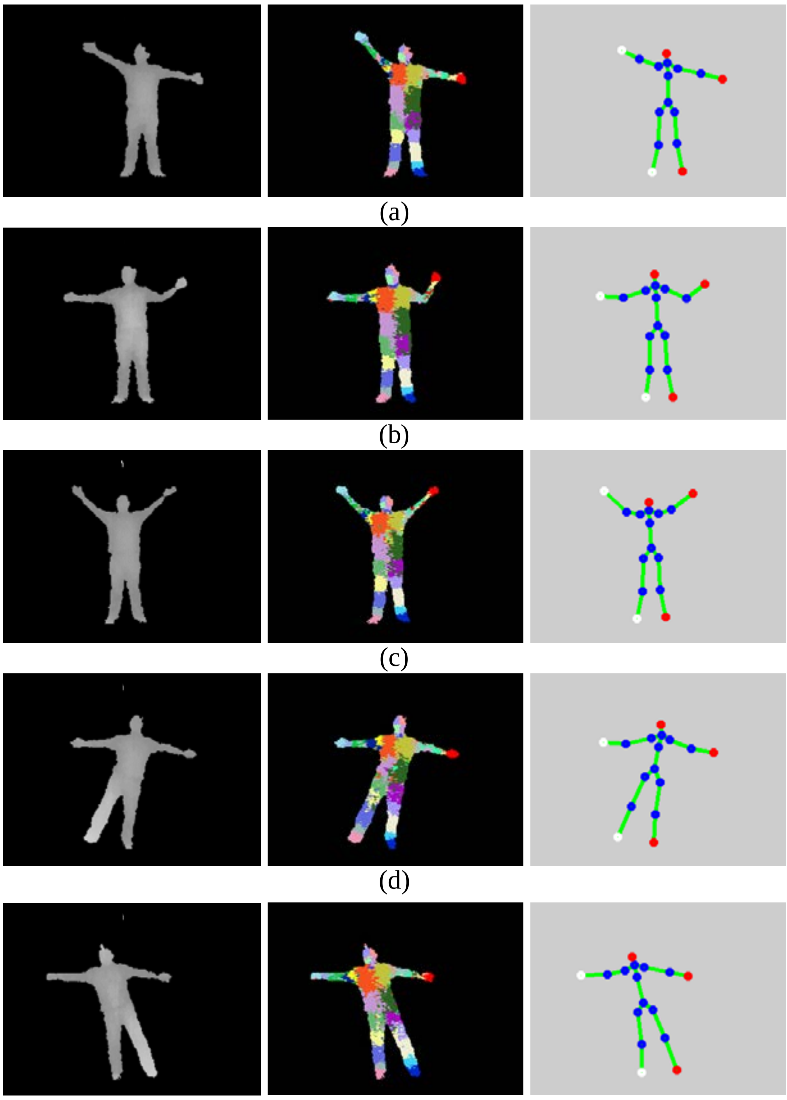


Fig. 4. A RF structure to train the labels of the depth silhouettes.



(e)

Fig. 5. A sample labeled body parts and corresponding joints from (a) right hand waving and (b) left ight hand waving, (c) both hand waving, (d) right leg movinh, and (e) left leg moving activity.

3.2 Body Segmentation

Random Forests (RFs) are an effective and reliable tool to deal with multi-class classification problems [14]. A forest is a combination of decision trees where each tree has nodes and leaves as shown in **Fig. 4**. In the figure, P represents the probability of M classes (i.e., labels here) using corresponding tree. For training RFs in HAR, first of all, simple features are built for each pixel in a depth image based on differences between neighboring pixel pairs. Thus, all features of all depth pixels and their corresponding labels obtained from training activity images are collected and used to train RFs, which are later used to label each pixel in the depth image of testing activity. This approach makes the body part labeling approach very fast.

To create trained RFs, an ensemble of three decision trees is used where the maximum depth of each tree is 20. Each tree in the RFs is trained with different pixels sampled randomly from the training depth silhouettes and their corresponding body part labeled. A subset of two thousand training sample pixels is drawn randomly from each depth silhouette in the training activity image database. Final decision to label each depth pixel for a specific body part is based on voting of all trees in the RF. Finally, from this segmented body parts in each image, skeleton model representing 16 body joints is generated considering the labels of the body parts. **Fig. 5** shows a sample segmented body parts with different colors and joints from both hand waving and right hand waving activities. Thus, using RF for body segmentation based on random features, we can obtain a position and scale invariant human body skeleton in human activity videos.

3.3 Feature Extraction

As aforementioned, once the segmented depth body shape is available, a skeleton model representing 16 joints is obtained where each joint is denoted as Q. The body joints are considered and represented as head, neck, left shoulder, right shoulder, chest, central hip, left hip, right hip, right elbow, right palm, left elbow, left palm, left knee, right knee, left foot, and right foot respectively. Motion features representing motion parameters i.e., magnitude as well as direction of the joints in the next frame are computed for 16 joints. The magnitude T of a joint from two consecutive depth frames is as

$$T = \sqrt{(Q_{x(i-1)} - Q_{x(i)})^2 + (Q_{y_{(i-1)}} - Q_{y_{(i)}})^2 + (Q_{z(i-1)} - Q_{z(i)})^2}. \tag{6}$$

Thus, the size of 16 joint’s magnitude feature of each frame becomes a vector of 1x16. The angles of the same body joint between two consecutive frames are computed as

$$G_{Q(x,y)} = \tan^{-1}\left(\frac{Q_{y(i-1)} - Q_{y(i)}}{Q_{x(i-1)} - Q_{x(i)}}\right), \tag{7}$$

$$G_{Q(y,z)} = \tan^{-1}\left(\frac{Q_{z(i-1)} - Q_{z(i)}}{Q_{y(i-1)} - Q_{y(i)}}\right), \tag{8}$$

$$G_{Q(x.z)} = \tan^{-1}\left(\frac{Q_{z(i-1)} - Q_{z(i)}}{Q_{x(i-1)} - Q_{x(i)}}\right). \tag{9}$$

The size of the directional angles of the joints’ becomes a vector of 1x48. Hence, the motion features for each depth shape becomes with the size 1x64 altogether. The feature vector for a video frame is represented as F. Then, the features of the frames are augmented to represent N dimensional large features for corresponding video to represent F. All features of all videos are then used as input to a DBN for training, which is later on used for testing the motion features from an unknown activity video.

3.4 Deep Belief Network for Activity Modeling

Training a DBN consists of two key parts: namely pre-training and fine-tune. The pre-training phase is based on Bolt Restricted Boltzmann Machine (RBM). Once the network is pre-trained, weights of the networks are adjusted by fine-tune algorithm. RBM is useful for unsupervised learning and hence can contribute for avoiding local optimum errors. As shown in **Fig. 6**, two hidden layers are used for RBM. RBM is basically used to initialize the networks by unsupervised learning. In the initialization of the network, a greedy layer-wise training methodology is used. Once the weights of the first RBMs are trained, first hidden layer weights h_1 get fixed. Then, the weights of the second RBMs are trained using the previous hidden layer's fixed weights. At last, the output layer's RBMs are trained using the weights of second hidden layer weights h_2. For updating the weights, a contrastive divergence algorithm is used. When the weights are adjusted, a classic back propagation algorithm is utilized for adjusting all parameters.

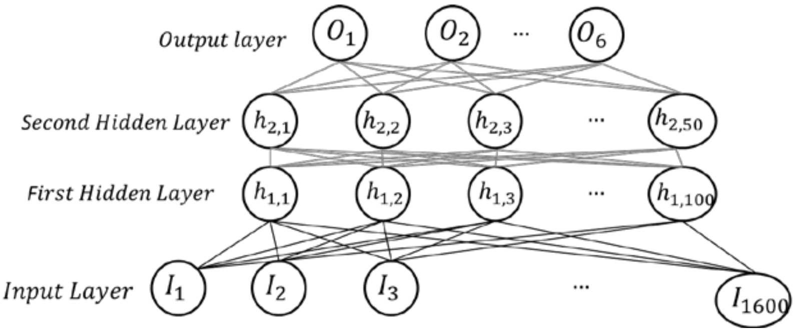


Fig. 6. Structure of a DBN used in this work.

4. Experimental Results

A human activity database of six different activities was built for experimental analysis. The activities were right leg moving, left leg moving, right hand waving, both hand waving, standing-up, and sitting-down. Hundred clips from each activity were collected to use for training. Finally, 100 clips were used to test each activity. Each video clip consisted of 26 frames. The experiments were started with the traditional binary and depth shape-based HAR with HMM. Since the binary silhouettes represent limited black and white color representations, the recognizer produced very poor recognition rates as shown in **Table 1** where the maximum average recognition rate was 77.33% using ICA approach. Basically, ICA represents better feature than PCA and the binary shape-based experiments also reflect that by achieving better performance using ICA than PCA. The experiments were then continued to the depth shape-based HAR and **Table 2** shows the experimental results where it shows the superiority of the depth shapes over the binary ones.

Finally, the body joint motion feature-based experiments were done where much better recognition performance than the binary as well as depth shape-based experiments was obtained as included in **Table 3**. Firstly, motion features were combined with HMM which achieved 91.33% recognition performance. For each HMM, we applied six states ergodic model. Later on, proposed approach (i.e., motion features with DBN) was tried that showed the highest recognition rate (i.e., 96.67%) over all other approaches and hence showing its superiority over others. The DBN consists of 1600 input neurons (i.e., 64x25), 100 neorons in hidden layer1, 50 neurons in hidden layer2, and finally 6 neurons (i.e., six activtiies) for output layer. The experiments were tried on a system with configuration as Intel® Core$^{(TM)}$ i3 CPU (2 cores and 4 logical processors), 8 GB RAM, Windows 8 Pro Operating system, and Matlab 2015a. We tried different approaches for training and testing multiple times. **Table 4** shows the average training time of features of all frames of all videos of all activities and average testing time for features of a single video. The highest training time was required by ICA (i.e., 109.23 seconds) as IC extraction from the all training depth silhouttes was computationally expensive but testing was much fast (i.e., 0.022 second per video) as it required only IC feature weight matrix multiplication. The second highest training time was taken by DBN i.e., 23.11 seconds but testing with DBN was also very fast (i.e., 0.021 second per video) as the weights in the units of different layers were already adjusted during training. It can be noticed in the table that all approaches can be implemented in real time as their testing time for each video is really fast with the aformentioned computer configuration. Once the frames are acquired from the depth camera, the fast testing time of the activity features for each video via DBN indicates the implementaibiliy of the proposed system in real time very well.

Table 1. HAR experimental on binary silhouettes results using different approaches.

Approach	Actvitiy	Recognition Rate	Mean
PCA-HMM	Right Leg Moving	67	**70.67**
	Left Leg Moving	69	
	Right Hand Waving	74	
	Both Hand Waving	73	
	Sitting-Down	76	
	Standing-Up	67	
ICA-HMM	Right Leg Moving	79	**77.33**
	Left Leg Moving	74	
	Right Hand Waving	81	
	Both Hand Waving	73	
	Sitting-Down	79	
	Standing-Up	83	

Table 2. HAR experimental on depth silhouettes results using different approaches.

Approach	Actvitiy	Recognition Rate	Mean
PCA-HMM	Right Leg Moving	75	**76.67**
	Left Leg Moving	77	
	Right Hand Waving	83	
	Both Hand Waving	81	
	Sitting-Down	71	
	Standing-Up	73	
	Right Leg Moving	87	
	Left Leg Moving	85	

	Right Hand Waving	91	**87.33**
	Both Hand Waving	89	
	Sitting-Down	83	
	Standing-Up	89	

Table 3. HAR experimental results on body joint motion features using different approaches.

Approach	**Actvitiy**	**Recognition Rate**	**Mean**
Body Joint Motion Feature-based HAR with HMM	Right Leg Moving	93	**91.33**
	Left Leg Moving	88	
	Right Hand Waving	95	
	Both Hand Waving	91	
	Sitting-Down	89	
	Standing-Up	94	
Body Joint Motion Feature-based HAaturR with DBN	Right Leg Moving	95	**96.67**
	Left Leg Moving	93	
	Right Hand Waving	99	
	Both Hand Waving	95	
	Sitting-Down	97	
	Standing-Up	96	

Table 4. Average training and testing time for different HAR approaches.

Approach	**Average Training Time of Features of All Frames of All Videos (Seconds)**	**Average Testing Time of Features of Frames of a Sngle Video (Seconds)**
PCA-HMM	5.02	0.015
ICA-HMM	109.23	0.022
Body Joint Features with HMM	3.26	0.018
Body Joint Features with DBN	23.11	0.021

4.1 Experiments on MSRDailyActivity3D Dataset

We also tried our approach on a public dataset named MSRDailyActivity3D dataset [29] that consisted of 16 daily activities: namely drink, eat, read book, call on cellphone, write on a paper, use laptop, use vacuum cleaner, cheer up, sit still, toss paper, play game, lie down on sofa, walk, play guitar, stand up, and sit down. The database had a total of 320 videos for which 10 subjects were involved. We tried a cross-subject testing/training for our experiments. **Table 5** shows the recognition results of the proposed approach where 91.56 % mean recognition rate was obtained. We considered 23 frames from each video. So 22 pair of consecutive frames are considered to extract features from each video. The DBN consisted of 1408 input neurons (i.e., 64x22), 100 neorons in hidden layer1, 50 neurons in hidden layer2, and finally 16 neurons (i.e., sixteen activtiies) for output layer. The proposed method was compared with other state-of-art methods where it obtained a superior performance over them as shown in **Table 6**. We tried the proposed approach with DBN for training and testing ten times where the average training time of features from all frames of all videos of all activities was 61.24 seconds. The average testing time for features from each video was 0.0.027 second, indicating really fast feature testing.

Table 5. HAR-experiment results for proposed-approach MSRDailyActivity3D dataset

Activity	Recognition Rate	Mean
Drink	90 %	**91.56**
Eat	90	
Read book	90	
Call on cell phone	95	
Write on a paper	90	
Use laptop	85	
Use vacuum cleaner	95	
Cheer up	95	
Sit still	95	
Toss paper	90	
Play game	85	
Lie down on sofa	90	
Walk	90	
Play guitar	95	
Stand up	95	
Sit down	95	

Table 6. Comparison of HAR performances of different approaches on MSRDailyActivity3D dataset.

Method	Recognition Accuracy
Wang et al. [31]	68.0 %
Dollar et al. [32]	73.6
Laptev et al. [33]	79.1
Lu and Aggarwal [34]	83.6
Cho et al. [1]	89.7
Our Proposed Approach	91.56

4.2 Experiments on MSRC-12 Gesture Dataset

Our HAR approach was also checked on MSRC-12 gesture dataset [30] where the dataset consisted of sequences of human skeletal joint movements and it has 594 sequences collected from 30 people for twelve different activities. There were 6244 gesture samples altogether. The activities were Lift arms, Duck, Push right, Goggles, Wind it up, Shoot, Bow, Throw, Had enough, Change weapon, Beat both, and Kick. We considered 26 frames from each video. So 25 pair of consecutive frames were considered to extract features from each video. We compared our deep learning on body joint motion feature-based approach with the traditional

HMM-based one where the proposed one showed the better accuracy (i.e., 97.93% mean) than traditional one (92.49% mean) as represented in **Table 7** and **Table 8**. For each HMM, we applied six states ergodic model. The DBN used in thie regard consists of 1600 input neurons (i.e., 64x25), 100 neurons in first hidden layer, 50 neurons in second hidden layer, and 12 neurons (i.e., twelve activities) for output layer. The proposed method obtained a superior performance over other state-of-art methods on the same dataset as shown in **Table 9**. We tried the proposed HAR approach with DBN for training and testing ten times with the aforementoined computer configuration. The average training time of features from all frames was 109.12. seconds. The mean testing time for features of each activity video was 0.031 second, showed quite fast feature testing.

Table 7. HAR-experiment results using traditional HMM-based approach on MSRC-12 dataset

Activity	**Recognition Rate**	**Mean**
Lift arms	87.5%	**92.49**
Duck	98.8	
Push right	84.2	
Goggles	91.8	
Wind it up	86.1	
Shoot	97.6	
Bow	92.9	
Throw	95.1	
Had enough	93.9	
Change weapon	94.8	
Beat both	92.4	
Kick	94.9	

Table 8. HAR-experiment results using proposed DBN-based approach on MSRC-12 dataset

Activity	**Recognition Rate**	**Mean**
Lift arms	96.1%	**97.93**
Duck	98.8	
Push right	98.8	
Goggles	96.9	
Wind it up	98.7	
Shoot	100	
Bow	98.8	
Throw	96.9	
Had enough	98.0	
Change weapon	96.1	
Beat both	98.7	
Kick	97.4	

Table 9. Comparison of HAR performances of different approaches on MSRC-12 dataset.

Method	Recognition Accuracy
HGM [35]	66.25%
ELC-KSVD [36]	90.22
Cov3DJ [37]	91.70
Our Proposed Approach	97.93

5. Conclusion

The goal of assisted living is to develop methods to promote the ageing in place of elderly people. Human activity recognition systems can help to monitor aged people in home environments. For this task, different sensors can be used. Among which, RGBD sensors seem to cost-effective and can provide much visual information about the environment. Our work aims to propose a novel human activity recognition method using motion features from skeleton data extracted by RGBD sensors and deep learning for modeling activities. The robust motion features of body joints are extracted through segmentation of different depth body parts using random forests. Then, DBN is used for activity learning and recognition. The experimental results on our dataset showed quite significantly improved recognition performance (i.e., 96.67%) using proposed approach than the conventional approaches (i.e., 91.33% at best). The proposed deep learning-based approach was also applied on MSRDailyActivity3D and MSRC-12 public datasets where it showed superior performance on some state-of-the-art methods by achieving mean recognition rate of 91.56% and 97.93% respectively. The proposed HAR system can be effectively employed to many smart applications such as smart home healthcare to monitor human activities in a smart home which can contribute to improve the quality of a user's life. In future, we aim to consider the occluded human body regions in complex human activities and body part segmentations for activity postures to extract missing skeleton joints in occlusion. It should make our human activity recognition dynamically applicable in real time smart environments.

References

[1] S.-S. Cho, A-Reum Lee, H.-I. Suk, J.-S. Park, and S.-W. Lee, "Volumetric spatial feature representation for view invariant human action recognition using a depth camera," *Optical Engineering*, vol. 54(3), no. 033102, pp. 1-8, 2015. Article (CrossRef Link)

[2] A. Jalal ., M.Z. Uddin, J.T. Kim, and T.S. Kim, "Recognition of human home activities via depth silhouettes and R transformation for smart homes," *Indoor and Built Environment*, vol. 21, no. 1, pp. 184-190, 2011. Article (CrossRef Link)

[3] M. Z. Uddin and T.-S. Kim, "Independent Shape Component-based Human Activity Recognition via Hidden Markov Model," *Applied Intelligence*, vol. 2, pp. 193-206, 2010. Article (CrossRef Link)

[4] N. Robertson and I. Reid, "A General Method for Human Activity Recognition in Video," *Computer Vision and Image Understanding*, Vol. 104, No. 2. pp. 232 – 248, 2006. Article (CrossRef Link)

[5] H. Kang, C. W. Lee, and K. Jung, "Recognition-based gesture spotting in video games," *Pattern Recognition Letters*, Vol. 25, pp. 1701-1714, 2004. Article (CrossRef Link)

[6] F.S. Chen, C.M. Fu, and C.L. Huang, "Hand gesture recognition using a real-time tracking method and Hidden Markov Models," *Image and Vision Computing*, vol. 21. pp.745-758, 2005. Article (CrossRef Link)

[7] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden markov model," in *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 379-385, 1992. Article (CrossRef Link)

[8] F. Niu and M. Abdel-Mottaleb, "View-invariant human activity recognition based on shape and motion Features," in *Proc. of IEEE Sixth International Symposium on Multimedia Software Engineering*, pp. 546-556, 2004. Article (CrossRef Link)

[9] F. Niu and M. Abdel-Mottaleb, "HMM-based segmentation and recognition of human activities from video sequences," in *Proc. of IEEE International Conference on Multimedia & Expo.* , pp. 804-807, 2005. Article (CrossRef Link)

[10] P. Simari, D. Nowrouzezahrai, E. Kalogerakis, and K. Singh, "Multi-objective shape segmentation and labeling," in *Proc. of Eurographics Symposium on Geometry Processing*, Vol. 28, pp. 1415-1425, 2009. Article (CrossRef Link)

[11] V. Ferrari, M.-M. Jimenez, and A. Zisserman, "2D Human Pose Estimation in TV Shows," *Visual Motion Analysis, LNCS 2009*, Vol. 5604, pp. 128-147, 2009. Article (CrossRef Link)

[12] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "A Full-Body Layered Deformable Model for Automatic Model-Based Gait Recognition," *EURASIP Journal on Advances in Signal Processing*, Vol. 1, pp. 1-13, 2008. Article (CrossRef Link)

[13] J. Wright and G. Hua, "Implicit Elastic Matching with Random Projections for Pose-Variant face recognition," in *Proc. of IEEE conf. on Computer Vision and Pattern Recognition*, pp. 1502-1509, 2009. Article (CrossRef Link)

[14] A. Bosch, A. Zisserman, and X. Munoz, "Image classification using random forests and ferns," in *Proc. of IEEE Int. Conf. on Computer Vision* , pp. 1-8, 2007. Article (CrossRef Link)

[15] P. S. Aleksic and A. K. Katsaggelos, "Automatic facial expression recognition using facial animation parameters and multistream HMMs," *IEEE Transaction on Information and Security*, vol. 1, pp. 3-11, 2006. Article (CrossRef Link)

[16] M. Z. Uddin and M. M. Hassan, "A Depth Video-Based Facial Expression Recognition System Using Radon Transform, Generalized Discriminant Analysis, and Hidden Markov Model," *Multimedia Tools And Applications*, Vol. 74, No. 11, pp. 3675-3690, 2015. Article (CrossRef Link)

[17] W Li., Z. Zhang, and. Z. Liu, "Action recognition based on a bag of 3d points," in *Proc. of Workshop on Human Activity Understanding from 3D Data*, pp. 9-14, 2010. Article (CrossRef Link)

[18] O. Oreifej and Z. Liu, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 716-723, 2013. Article (CrossRef Link)

[19] X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion mapsbased histograms of oriented gradients," in *Proc. of ACM International Conference on Multimedia*, pp. 1057-1060, 2012. Article (CrossRef Link)

[20] H.S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from RGB-D videos," *International Journal of Robotics Research*, vol. 32, no. 8, pp. 951-970, 2013. Article (CrossRef Link)

[21] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Unstructured human activity detection from rgbd images," in *Proc. of IEEE International Conference on Robotics and Automation*, pp. 842-849, 2012. Article (CrossRef Link)

[22] H. Hamer, K. Schindler, E. Koller-Meier, and L. Van Gool, "Tracking a hand manipulating an object," in *Proc. of IEEE International Conference on Computer Vision*, pp. 1475-1482, 2009. Article (CrossRef Link)

[23] P. Dreuw, H. Ney, G. Martinez, O. Crasborn, J. Piater, J.M. Moya, and M. Wheatley, "The signspeak project - bridging the gap between signers and speakers," in *Proc. of International Conference on Language Resources and Evaluation*, pp. 476-481, 2010. Article (CrossRef Link)

[24] M.Z. Uddin, N.D. Thang, and T.S. Kim, "Human Activity Recognition Using Body Joint Angle Features and Hidden Markov Model," *ETRI Journal*, pp. 569-579, 2011. Article (CrossRef Link)

[25] G. Hinton, L. Deng, D. Yu, G.E. Dahl, A.R. Mohamed, N. Jaitly, A. Vanhoucke, Nguyen, P., T.N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp.82-97, 2012. Article (CrossRef Link)

[26] G. E. Hinton, S. Osindero, Y. The, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, pp. 1527-1554, 2006. Article (CrossRef Link)

[27] S. Izadi, "KinectFusion: Real-time 3D Reconstruction and Interaction Using a Moving Depth Camera," in *Proc. of ACM User Interface and Software Technologies*, pp. 559-568, 2011. Article (CrossRef Link)

[28] Y. m. Song, S. Noh, J. Yu, C. w. Park, B. g. Lee, "Background subtraction based on Gaussian mixture models using color and depth information," in *Proc. of International Conference on Control, Automation and Information Sciences (ICCAIS)*, pp. 132 - 135, 2014. Article (CrossRef Link)

[29] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. of 2012 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 1290–1297, 2012. Article (CrossRef Link)

[30] S. Fothergill, H. Mentis, P. Kohli, and S. Nowozin, "Instructing people for training gestural interactive systems," in *Proc. of ACM Conference on Human Factors in Computing Systems*, pp. 1737-1746, 2012. Article (CrossRef Link)

[31] J. Wang, Z. Liu, Y. Wu , and Junsong Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. of IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 1290–1297, IEEE, Providence, 2012. Article (CrossRef Link)

[32] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. of 2nd Joint IEEE Int. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 65–72, IEEE, Washington, 2005. Article (CrossRef Link)

[33] I. Laptev, R. Rennes, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, Anchorage, 2008. Article (CrossRef Link)

[34] X. Lu and J. Aggarwal, "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera," in *Proc. of IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 2834–2841, IEEE, Portland , 2013. Article (CrossRef Link)

[35] S. Yang, C. Yuan, W. Hu, and X. Ding, "A hierarchical model based on latent dirichlet allocation for action recognition," in *Proc. of IEEE 22nd International Conference on In Pattern Recognition (ICPR)*, pp. 2613–2618. 2014. Article (CrossRef Link)

[36] L. Zhou, W. Li, Y. Zhang, P. Ogunbona, D. T. Nguyen, and H. Zhang, "Discriminative key pose extraction using extended lc-ksvd for action recognition," in *Proc. of International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–8. IEEE, 2014. Article (CrossRef Link)

[37] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. ElSaban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations," in *Proc. of International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2466–2472, 2013. Article (CrossRef Link)

Md. Zia Uddin got his Ph.D. degree in Biomedical Engineering from Kyung Hee University of South Korea in February of 2011. Now, he is working as a post-doctoral research fellow under Robotics and Intelligent Systems (ROBIN) research group at Dept. of Informatics, University of Oslo, Norway. Dr. Zia's researches are mainly focused on computer vision, image processing, and pattern recognition. More specifically, he has worked on colour as well as 3-D depth video-based human activity, gait, and facial expression analysis. In this regard, he has developed several novel methodologies using colour and depth videos. His works offer many practical applications in the areas of smart homes, human computer interfaces, life-care, healthcare, etc. For instance, in a smart home, a human activity recognition system can recognize its resident's activities automatically and can create daily, monthly, and yearly activity databases. For video games, a user's action can be recognized without a controller or attached sensors. His research outcomes have been published in famous journals such as IEEE transactions on consumer electronics, etc. He got more than 60 research publications including international journals, conferences, and book chapters.

Professor Jaehyoun Kim received his B.S. degree in mathematics from Sungkyunkwan University, Seoul, Korea, M.S. degree in computer science from Western Illinois University and Ph.D. degrees in computer science from Illinois Institute of Technology in U.S.A. He was a Chief Technology Officer at Kookmin Bank in Korea before he joined the Department of Computer Education at Sungkyunkwan University in March 2002. Currently he is a professor at Sungkyunkwan University. His research interests include software engineering & architecture, e-Learning, SNS & communication, internet business related policy and computer based learning.