Human Activity Recognition Using Spatiotemporal 3-D Body Joint Features with Hidden Markov Models

Md. Zia Uddin and Jaehyoun Kim
Department of Computer Education, Sungkyunkwan University
Seoul, South Korea
[e-mail: ziauddin@skku.edu, jaekim@skku.edu]
*Corresponding author: Jaehyoun Kim

Received June 24, 2015; revised December 4, 2015; accepted May 1, 2016; published June 30, 2016

Abstract

Video-based human-activity recognition has become increasingly popular due to the prominent corresponding applications in a variety of fields such as computer vision, image processing, smart-home healthcare, and human–computer interactions. The essential goals of a video-based activity-recognition system include the provision of behavior-based information to enable functionality that proactively assists a person with his/her tasks. The target of this work is the development of a novel approach for human-activity recognition, whereby human-body-joint features that are extracted from depth videos are used. From silhouette images taken at every depth, the direction and magnitude features are first obtained from each connected body-joint pair so that they can be augmented later with motion direction, as well as with the magnitude features of each joint in the next frame. A generalized discriminant analysis (GDA) is applied to make the spatiotemporal features more robust, followed by the feeding of the time-sequence features into a Hidden Markov Model (HMM) for the training of each activity. Lastly, all of the trained-activity HMMs are used for depth-video activity recognition.

Keywords: Human-activity Recognition, Body Joints, Generalized Discriminant Aanalysis, Hidden Markov Models

This work was supported by the Samsung Research Fund, Sungkyunkwan University, 2015.

http://dx.doi.org/10.3837/tiis.2016.06.017 **ISSN : 1976-7277**

1. Introduction

The recognition of a variety of human activities from video has emerged as a key research area in computer vision, image processing, and human–computer interaction (HCI). Over the last decade, human activity recognition (HAR) systems have therefore attracted a great deal of attention from a community of respected researchers due to their applications in many areas of pattern recognition and computer vision [1]–[11]. The accurate recognition of human activities, however, is still considered a major concern for most of these researchers due to the lack of accuracy that can occur because of a variety of causes such as a failed efficient-feature extraction, a low variance among the features of different activities, and a high variance among the features of the same activity class.

1.1 Related Works

For video-based HAR, the use of binary silhouettes is considered the most-popular approach [1]–[11]; for instance, in [4], where binary pixel-based mesh features were extracted from every image, the authors used binary silhouettes for HAR. In [5] and [6], the authors adopted Principal Component (PC)-based binary silhouette features to recognize view-invariant human activities. In [8], the authors proposed the use of the Independent Component (IC) features of binary silhouettes to recognize five different activities by means of a Hidden Markov Model (HMM), and they showed the superiority of the IC-based local features over the PC-based global-silhouette features. Although binary silhouettes are commonly employed to represent a wide variety of body configurations, they cannot be used to distinguish far and near body parts; therefore, binary silhouettes are clearly not suitable for the representation of different postures. To improve body silhouette representation, one could utilize depth information like the authors of [9] who derived IC features from time-sequence-activity depth silhouettes for a robust HAR. In [10], the authors applied a linear discriminant analysis (LDA) to represent robust activity-silhouette features. Basically, the functionality of an LDA is based on the class information that projects data onto a subspace by using the criterion that tries to maximize the between-class scatterings and minimize the within-class scatterings of the projected data. A generalized discriminant analysis (GDA) [11] that tries to separate the class samples using nonlinear subspace, however, can be preferable to an LDA in terms of its applicability regarding activity features; therefore, a GDA can be considered a robust tool for the classification of human-activity features.

Although depth silhouettes seem to be more effective than binary silhouettes, some ambiguities such as body-part segmentation remain unresolved. Since human-body parts are connected, body-joint features can lead to a robust HAR when compared to a HAR for which whole-body features are used. Human-body-joint analysis has received a great deal of attention from many computer-vision researchers [12]–[16]. In [12], k-means clustering was applied for the segmentation of body silhouettes to obtain body joints. In [13], the authors proposed an upper-body-part analysis for the representation of a person in a human-pose estimation. In [14], the authors applied a manual framework to segment human-body silhouettes so that the body joints could be obtained for gait recognition. Depth-information-based pattern analysis has attracted many researchers regarding a variety of applications such as human-motion analysis [17]–[38]. In [17], the authors used depth-map sequences for a human-activity analysis. In [19], the authors presented a depth-video-based human-activity analysis for which surface-orientation histograms were used. In [22], the

authors applied Depth Motion Maps (DMM) to capture the motion energies in activity videos. In [26], the authors focused on joint activity and object labeling from RGB and depth videos for an activity analysis. In [28], the authors represented human activities using a two-layer Maximum Entropy Markov Model (MEMM). In [32], the authors used particle swarm optimization (PSO) to model two interacting hands from depth images. For depth-information-based works, visual gestural languages such as American Sign Language (ASL) also form a very active field in computer visions [35]–[38]; for example, the authors of the Sign Speak project [38] analyzed textual representations of continuous visual sign language. In [39], the authors first extracted the angle-based spatial features of body joints from noisy depth images obtained by stereo cameras that were then further applied with HMMs for different-activity recognition. While human-activity videos represent time-sequence events, spatiotemporal features can describe human activities in video more effectively than spatial features. As proposed in this work, it is therefore possible to derive a robust HAR by applying body-joint spatiotemporal features with a nonlinear feature-classification technique such as a GDA that can be further modelled by HMMs.

1.2 Proposed HAR Approach

In this work, a novel HAR approach is proposed whereby human-body-joint features are utilized with a GDA and HMMs. The connected body-joint magnitude and directional features are first generated from each depth body silhouette before they are augmented with the magnitude; this is followed by the application of the directional motion features of the joints in the next frame with the GDA to increase the feature robustness. The feature sequences are then applied to train each activity HMM so that they can be used later for an activity recognition that is based on the maximum likelihood. The proposed system consists of depth-video acquisition, activity-feature generation, and HMMs.

2. Proposed HAR Methodology

The proposed HAR system consists of depth-video acquisition, activity-feature generation, and modeling-activity HMMs. **Fig. 1** shows the architecture of the proposed HAR system.

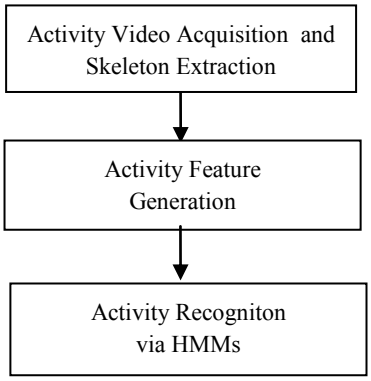

Fig. 1. Architecture of proposed depth-sensor-based HAR system.

2.1 Activity-video Acquisition and Skeleton Extraction

Kinect, a commercial depth-sensor-based camera, is utilized in this work to acquire the RGB as well as the depth images of different human activities [40]. The depth silhouette is then extracted from each depth image after a background subtraction. After obtaining a depth silhouette, a corresponding skeleton-body model that provides 15 joint positions is obtained through the OpenNI library [41]; therefore, from each depth silhouette, 15 body joints are obtained for the calculation of the spatiotemporal features. **Fig. 2 (a), (b), (c), and (d)** show a sample body-activity silhouette and the corresponding joints of both-hand waving, walking, running, and sitting activity, respectively. Each skeletal 3-D joint is represented as (D_x, D_y) with a depth value D_z.

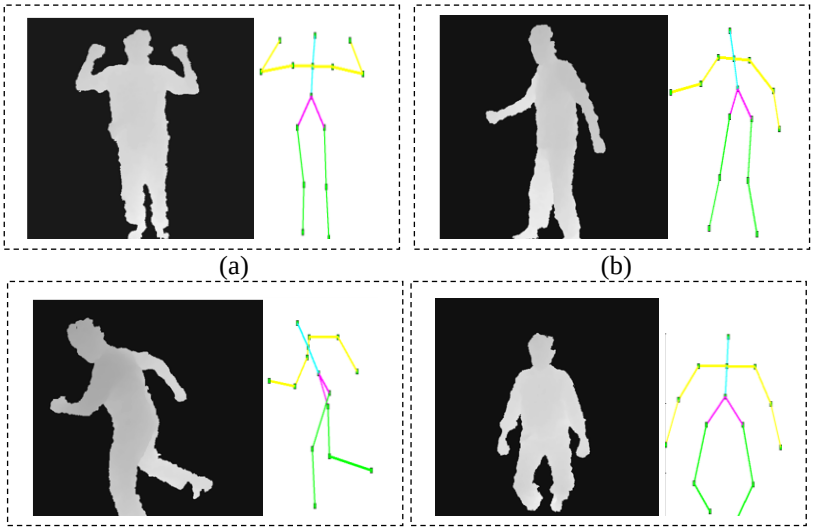

Fig. 2. Sample body-activity silhouette and the corresponding skeletal joints of (a) both-hand waving, (b) walking, (c) running, and (d) sitting activity.

2.2 Activity-feature Generation

After obtaining a depth silhouette, a corresponding skeleton-body model that provides 15 joint positions is obtained through the OpenNI library [41]; therefore, from each depth silhouette, 15 body joints are obtained for the calculation of the spatiotemporal features. The first feature information is the magnitude of the connected human-body joints; therefore, considering that each joint position in 3-D is (D_x, D_y) with a depth value D_z, the magnitude feature C of a connected joint pair can be expressed as the following:

$$C = \sqrt{(D_{x(i-1)} - D_{x(i)})^2 + (D_{y(i-1)} - D_{y(i)})^2 + (D_{z(i-1)} - D_{z(i)})^2}. \quad (1)$$

The size of the magnitude feature of each frame of 14 body-joint pairs becomes a vector of 1 x 14. The directional angles of each body-joint pair can be represented as the following:

$$A_{D(x,y)} = \arctan\left(\frac{D_{y(i-1)} - D_{y(i)}}{D_{x(i-1)} - D_{x(i)}}\right), \tag{2}$$

$$A_{D(y,z)} = \arctan\left(\frac{D_{z(i-1)} - D_{z(i)}}{D_{y(i-1)} - D_{y(i)}}\right), \tag{3}$$

$$A_{D(x,z)} = \arctan\left(\frac{D_{z(i-1)} - D_{z(i)}}{D_{x(i-1)} - D_{x(i)}}\right). \tag{4}$$

Later, the magnitude and directional features for each body joint, in consideration of its motion in the next frame, are calculated using (1)–(4); therefore, the temporal magnitude and directional-feature size of 15 body joints for a consecutive frame pair become a vector of 1 x 60. These motion features are then augmented with the connected-joint-pair spatial features to increase their robustness. The total size for a depth frame becomes 1 x 116 and can be denoted as F. **Fig. 3** shows a sample mean of the magnitude features from walking and running activity, respectively. The figure clearly shows that the motion magnitudes for running activity are higher than those for walking activity.

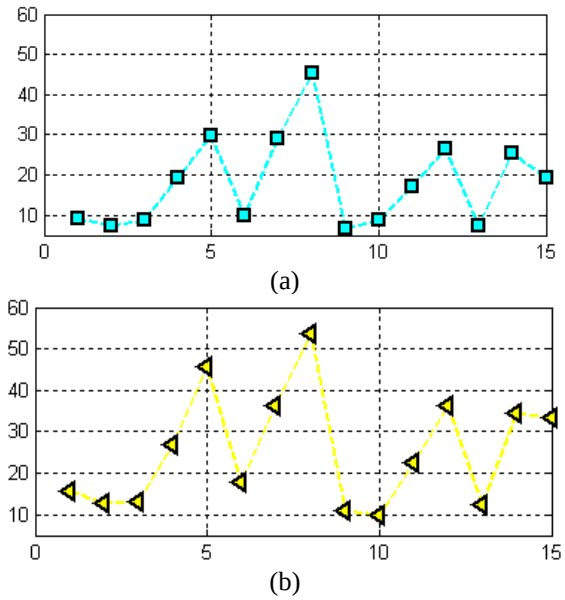

Fig. 3. Mean motion-parameter (magnitude) features from the joints of (a) walking- and (b) running-activity image sequences.

The final step in this regard is the application of the GDA. The GDA is based on class-specific information that maximizes the ratio of the within- T - and between- B -class

scatter matrices. Before determining the feature space, each depth-image feature is mapped into a Gaussian kernel. The optimized GDA discrimination matrix is determined as the following:

$$W_{GDA} = \arg \max_{LDA} \frac{\left|W^T BW\right|}{\left|W^T TW\right|}. \tag{5}$$

The final feature vectors can therefore be obtained, as follows:

$$M = FW_{GDA}^T. \tag{6}$$

where W_{GDA}^T is the optimal discrimination matrix that maximizes the ratio of T and B. After applying the GDA, the feature vector size for each depth silhouette becomes 1 x 5 due to the six different activity classes, indicating a huge dimension reduction, as well as a robust feature representation. **Fig. 4** shows an exemplar 3-D plot after the application of the GDA on the spatiotemporal features of six human activities where most of the samples from different classes are well clustered.

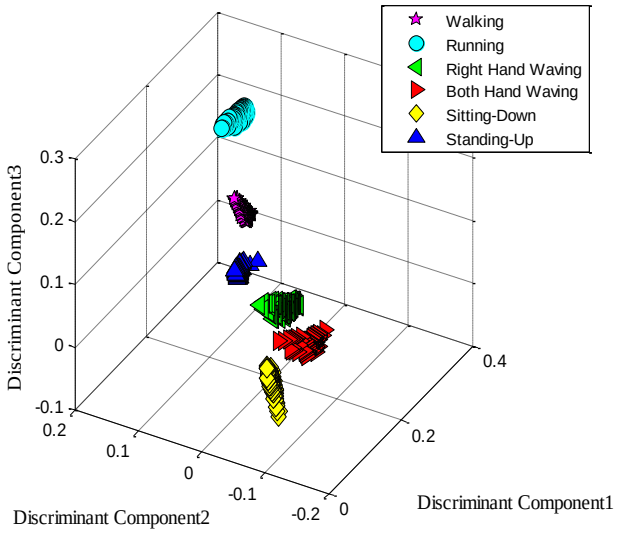

Fig. 4. Exemplar 3-D plot after application of GDA on the position and motion features of six human activities.

2.3 Activity Recognition via HMMS

For activity modeling and recognition in this work, we applied HMMs; essentially, HMMs are used for the handling of time-sequence features with a probabilistic learning process. We used an ergodic HMM structure to encode the time-sequence features. For the training HMMs, we applied the discrete HMMs with the codebook of 32. **Fig. 5** shows the basic steps for the generation of a codebook using the training body silhouettes and a symbol from a depth silhouette. **Fig. 6** shows a sample symbol sequence for an image sequence of six different

activities; in the figure, it is noticeable that the symbol sequences of the different activities are different from each other.

In the learning HMM, each HMM that corresponds to an activity is optimized by the discrete symbol sequences obtained from the training image sequences of that activity, whereby each activity is modeled by a trained HMM; therefore, for the K activities, a dictionary of K trained HMMs is created. To test an activity in a video, the corresponding observed symbol sequence is generated and applied on all of the trained HMMs to calculate the likelihood, and the one with the highest likelihood is selected. To test a sequence θ, the following activity decision is derived:

$$Decision = \arg\max_{i=1}^{K}(\mathrm{P}(\theta \mid H_i)). \tag{7}$$

More details regarding HMMs are available at [1], [4]–[6], and [8]–[11]. **Fig. 7** shows the structure of the ergodic HMM structure used in this work on our database, as the number of states is small and the number of states beyond four could not improve our dataset.

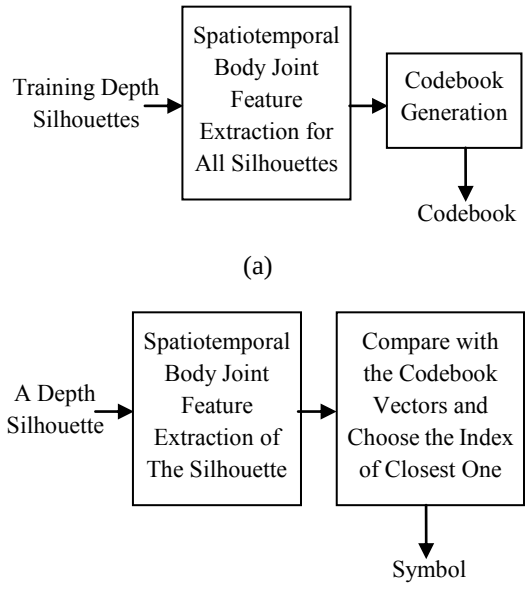

Fig. 5. Basic steps for the generation of (a) a codebook from the training silhouettes and (a) a symbol from a depth silhouette.

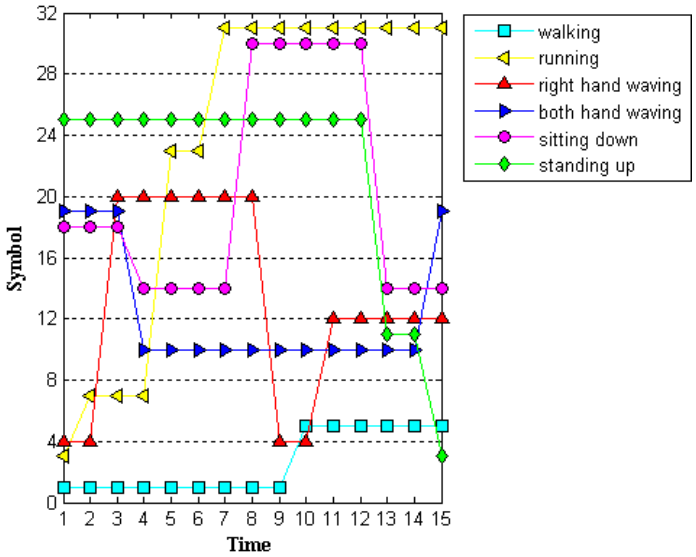

Fig. 6. A sample discrete symbol sequence for a sample image sequence of six different activities.

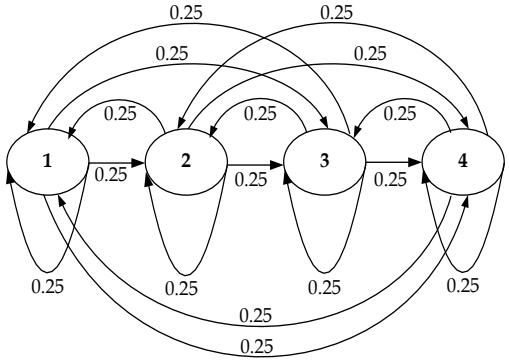

Fig. 7. Structure of a four-state ergodic HMM

3. HAR-experiment Results

A HAR database was built for the six different activities (walking, running, right-hand waving, both-hand waving, standing up, and sitting down) that were trained and recognized via the proposed approach. A total of 10 clips of variable lengths for each activity were used to build the training feature space, followed by the testing of 25 video clips for HAR. The experiments commenced with the depth-silhouette-based (i.e., without a 3-D body-joint basis) HAR first. Four different conventional feature-extraction methods—PCA, ICA, radon transformation, and ICA-GDA—were applied with the HMM to evaluate their performances in terms of depth-silhouette-based activity recognition. The PCA global feature-extraction method that was applied first achieved a 79.33 % recognition rate, indicating a poor HAR performance. As the ICA method represents activity-silhouette features that are more effective than those of the PCA method, the ICA method was applied, achieving an 88 % recognition rate that shows a HAR performance that is a great improvement upon that of the PCA method. The application of the radon-transformation feature-extraction method on the depth silhouettes achieved an 89.33 % mean recognition rate. Lastly, the extension of the ICA method using the GDA

method achieved the highest recognition performance of 90 %. **Table 1** shows the HAR recognition results of the PCA, ICA, radon-transformation, and ICA-GDA methods.

Table 1. HAR-experiment results for different depth-silhouette-based approaches

Approach	Activity	Recognition Rate	Mean
PCA on Depth Silhouettes	Walking	84.0 %	**79.33**
	Running	78	
	Right-hand Waving	88	
	Both-hand Waving	88	
	Sitting Down	74	
	Standing Up	64	
ICA on Depth Silhouettes	Walking	84	**88**
	Running	88	
	Right-hand Waving	92	
	Both-hand Waving	92	
	Sitting Down	88	
	Standing Up	84	
Radon Transformation on Depth Silhouettes	Walking	88	**89.33**
	Running	88	
	Right-hand Waving	92	
	Both-hand Waving	92	
	Sitting Down	88	
	Standing Up	88	
ICA-GDA on Depth Silhouettes	Walking	88	**90**
	Running	88	
	Right-hand Waving	92	
	Both-hand Waving	92	
	Sitting Down	92	
	Standing Up	88	

The experimental stage was continued to incorporate a spatiotemporal, 3-D body-joint, feature-based HAR. First, we extracted the body-joint-pair (limb)-based spatial features that were applied with the HMMs and obtained a mean recognition rate of 91.33 %. The spatial features were then augmented with the temporal (i.e., motion) features for each frame and applied with the HMMs; this approach achieved a 94 % recognition rate that is higher than those of all of the methods that have been mentioned so far. Lastly, the spatiotemporal features were enhanced by the GDA, achieving a superior recognition rate of 98.66 % that is the highest recognition performance. The proposed approach therefore shows a performance that is superior to the other HAR approaches. **Table 2** shows the spatiotemporal, feature-based HAR results.

Table 2. HAR-experiment results for 3-D body-joint-based approaches

Approach	Activity	Recognition Rate	Mean
Body-joint Spatial Feature-based HAR	Walking	92	**91.33**
	Running	88	
	Right-hand Waving	92	
	Both-hand Waving	92	
	Sitting Down	92	
	Standing Up	92	
Spatiotemporal Body-joint Feature-based HAR without GDA	Walking	92	**94**
	Running	92	
	Right-hand Waving	96	
	Both-hand Waving	96	
	Sitting Down	96	
	Standing Up	92	
Spatiotemporal Body-joint Feature-based HAR with GDA	Walking	96	**98.66**
	Running	96	
	Right-hand Waving	100	
	Both-hand Waving	100	
	Sitting Down	100	
	Standing Up	100	

Table 3. HAR-experiment results for proposed-approach MSRDailyActivity3D dataset

Activity	Recognition Rate	Mean
Drink	90 %	**90**
Eat	85	
Read book	90	
Call on cell phone	90	
Write on a paper	85	
Use laptop	85	
Use vacuum cleaner	95	
Cheer up	95	
Sit still	95	
Toss paper	85	
Play game	80	
Lie down on sofa	90	
Walk	85	
Play guitar	95	
Stand up	95	
Sit down	90	

3.1 Experiments on MSRDailyActivity3D Dataset

The MSRDailyActivity3D dataset [42] consists of daily activities that were captured by Microsoft Research using a Kinect device and comprises the following 16 activities: 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 consists of a total of 320 video samples for which 10 subjects were involved. A cross-subject training/testing setup was used for the experiments, and **Table 3** shows the recognition performance of the proposed approach regarding the MSRDailyActivity3D dataset, whereby a 90 % mean recognition rate was achieved. The proposed method was also compared with the state-of-art methods proposed in [42], [43], [44], and [45], and it showed a superior performance over the others, as shown in **Table 4**.

Table 4. Comparison of recognition performances for MSRDailyActivity3D dataset.

Method	Recognition Accuracy
Wang et al. [41]	68.0 %
Dollar et al. [43]	73.6
Laptev et al. [44]	79.1
Lu and Aggarwal [42]	83.6
Proposed Approach	90.0

4. Concluding Remarks

In this paper, a novel work has been proposed for human-activity recognition, whereby the spatiotemporal features from 3-D skeleton-body joints and HMMs are utilized. The proposed system was compared with the conventional approaches, whereby its superiority over the other approaches was shown by the attainment of the highest recognition rates on different databases. The proposed HAR system can be employed in numerous smart applications including smart-home healthcare for the monitoring of human activities in a smart home that can contribute to the improvement of the quality of people's lives. Regarding a future work, our aim is the consideration of the occluded human-body regions for complex human activities, so that missing skeleton joints can be extracted to make our system applicable in a variety of real-time smart environments.

References

[1] 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)

[2] 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)

[3] 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)

[4] 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)

[5] 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)
[6] 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)
[7] A. Agarwal and B. Triggs, "Recovering 3D human pose from monocular images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 44-58, 2006. Article (CrossRef Link)
[8] 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)
[9] M. Z. Uddin, D. H. Kim, J. T. Kim, and T.-S. Kim, "An Indoor Human Activity Recognition System for Smart Home Using Local Binary Pattern Features with Hidden Markov Models," *Indoor and Built Environment*, vol. 22, pp. 289-298, 2013. Article (CrossRef Link)
[10] 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)
[11] P. Yu, D. Xu, and P. Yu "Comparison of PCA, LDA and GDA for Palm print Verification," in *Proc. of International Conference on Information, Networking and Automation*, pp. 148-152, 2010. Article (CrossRef Link)
[12] P. Simari, D. Nowrouzezahrai, E. Kalogerakis, and K. Singh, "Multi-objective shape segmentation and labeling," *Eurographics Symposium on Geometry Processing*, Vol. 28, pp. 1415-1425, 2009. Article (CrossRef Link)
[13] 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)
[14] 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)
[15] 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)
[16] A. Bosch, A. Zisserman, and X. Munoz, "Image classification using random forests and ferns," *IEEE Int. Conf. on Computer Vision* , pp. 1-8, 2007. 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] W Li., Z. Zhang, and. Z. Liu, "Expandable data-driven graphical modeling of human actions based on salient postures," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 18, No. 11, pp. 1499-1510, 2008. Article (CrossRef Link)
[19] 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)
[20] A. Vieira, E. Nascimento, G. Oliveira, Z. Liu, and M. Campos, "Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences," in *Proc. of Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pp. 252-259, 2012. Article (CrossRef Link)
[21] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu, "Robust 3d action recognition with random occupancy patterns," in *Proc. of European Conference on Computer Vision*, pp. 872-885, 2012. Article (CrossRef Link)
[22] 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)

[23] J. Lei, X. Ren, and D. Fox, "Fine-grained kitchen activity recognition using rgb-d," in *Proc. of ACM Conference on Ubiquitous Computing*, pp.208-211, 2012. Article (CrossRef Link)
[24] M.Z. Uddin, T.S. Kim, and J.T. Kim, "A Spatiotemporal Robust Approach for Human Activity Recognition," *International Journal of Advanced Robotic Systems*, 2013. Article (CrossRef Link)
[25] Y. Wang, K. Huang, and T. Tan, "Human activity recognition based on r transform," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8, 2007. Article (CrossRef Link)
[26] 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)
[27] X. Yang and Y. Tian, "Eigenjoints-based action recognition using naive-bayesnearest-neighbor," in *Proc. of Workshop on Human Activity Understanding from 3D Data*, pp. 14-19, 2012. Article (CrossRef Link)
[28] 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)
[29] A. McCallum, D. Freitag, and F.C.N. Pereira, "Maximum entropy markov models for information extraction and segmentation," in *Proc. of International Conference on Machine Learning*, pp. 591-598,2000. Article (CrossRef Link)
[30] H. Hamer, K. Schindler, E. Koller-Meier, and L. Van Gool, "Tracking a hand manipulating an object," in *Proc. of International Conference on Computer Vision*, pp. 1475-1482, 2009. Article (CrossRef Link)
[31] I. Oikonomidis, N. Kyriazis, and A.A. Argyros, "Tracking the articulated motion of two strongly interacting hands," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1862-1869, 2012. Article (CrossRef Link)
[32] H. Hamer, J. Gall, T. Weise, and L. Van Gool, "An object-dependent hand pose prior from sparse training data," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition.* pp. 671-678, 2010. Article (CrossRef Link)
[33] D. D. Luong, S. Lee, and T.-S. Kim, "Human Computer Interface Using the Recognized Finger Parts of Hand Depth Silhouette via Random Forests," in *Proc. of 13th International Conference on Control, Automation and Systems*, pp. 905-909, 2013. Article (CrossRef Link)
[34] M. Z. Uddin, J. T. Kim, and T.-S. Kim, "Depth video-based gait recognition for smart home using local directional pattern features and hidden Markov model," *Indoor and Built Environment*, vol. 23, no. 1, pp.133-140, 2014. Article (CrossRef Link)
[35] S. Ong and S. Ranganath, "Automatic sign language analysis: A survey and the future beyond lexical meaning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 873-891, 2005. Article (CrossRef Link)
[36] T. Pei, T. Starner, H. Hamilton, I. Essa, and J. Rehg, "Learnung the basic units in american sign language using discriminative segmental feature selection," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4757-4760, 2009. Article (CrossRef Link)
[37] H.D. Yang, S. Sclaroff, and S.W. Lee, "Sign language spotting with a threshold model based on conditional random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no.7, pp.1264-1277, 2009. Article (CrossRef Link)
[38] 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)
[39] 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)
[40] 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)

[41] G. T. Papadopoulos, A. Axenopoulos, P. Daras, "Skeleton Tracking using Kinect Sensor & Displaying in 3D Virtual Scene," *International Journal of Advancements in Computing Technology*, vol. 4, no, 11, pp. 213-223, 2012. Article (CrossRef Link)
[42] 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)
[43] X. Lu and J. Aggarwal, "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera," in *Proc. of 2013 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 2834–2841, IEEE, Portland, 2013. Article (CrossRef Link)
[44] 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, 2005. Article (CrossRef Link)
[45] I. Laptev, 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, 2008. Article (CrossRef Link)

Md. Zia Uddin received his Ph.D. degree from Biomedical Engineering department of Kyung Hee University of South Korea in February 2011. He is working now as an Assistant Professor in Computer Education Department of Sungkyunkwan University (SKKU) of South Korea. SKKU is one of the leading universities in South Korea and it has good world ranking as well. Dr. Zia has got many research publications in international journals, conferences, and book chapters. His research interests include computer vision, image processing, smarthome healthcare, video surveillance, human computer interaction, human activity recognition, and facial expression recognition.

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.