A Two-Level Hidden Markov Model-based Approach for Human Activity Recognition

Md. Zia Uddin

Department of Computer Education

Sungkyunkwan University

Seoul, Republic of Korea

ziauddin@skku.edu

Abstract *:*

A novel approach is proposed in this work for human activity recognition from depth video sensor data using features of human body and two-level Hidden Markov Models. From depth video, different body parts of human activities are segmented first by means of random forests. Then, robust features are obtained using labeled body parts and body joint information which includes motion information as well. Traditionally, a dictionary of all trained HMMs for all the activities are built and the features are applied on the HMMs to generate the maximum likelihood to obtain the proper activity which is time consuming if the activities are many. Hence, two-level activity recognition is proposed in this work. Initially, all activities are distributed in some groups. In first level, activity group is determined by means of applying the features on activity class HMMs. Finally, the features are applied again on the trained activity HMMs of the group obtained from first level classification. The proposed approach shows better performance than traditional approaches.

Keywords-component; *Body Joint, Depth Information, Hidden Markov Models.*

I. INTRODUCTION

Video-based Human Activity Recognition (HAR) is a great research topic that has been attracting many researchers of computer vision nowadays [1]. For video-based HAR, 2-D flat binary silhouettes are most commonly applied [1]-[3]. In [1], the authors Principal Component Analysis (PCA) for representing the global features that result in very poor HAR performance. After PCA, Independent Component Analysis (ICA) has been applied for HAR. In [2], the authors used ICA-based local binary silhouette features for HAR that yielded better recognition performance than PCA-based ones. Despite of the easiness of implementation of binary silhouettes, they have limitations of distance information representation in the silhouette. However, depth information can overcome this problem where one can obtain the body silhouette pixel intensities using depth camera and utilize depth silhouettes for robust HAR [3]. Though depth silhouettes are really better than binary ones but it should be better if one can separate different body parts to get the joints in the image as the human body consists of body parts connected together. Hence, one can get robust features from body joints than considering whole body features for HAR.

Recently, body part labeling or segmentation is also getting pretty good attentions by image processing and computer vision researchers [4]-[8]. In [4], the authors applied k-means for body part segmentation and labeling.

In [5], the authors did upper body segmentation to estimate human body posture. In [6], the authors considered body part segmentation manually to obtain body joints to be applied for gait recognition.

In this work, a HAR novel approach is described using spatiotemporal features of body labels and joints with Hidden Markov Models (HMMs). For training activity, after obtaining the body labels and joints, spatiotemporal features of the body joints are generated from depth silhouettes in the activity videos. Then, the feature sequences from the activity videos are applied to train distinguished activity HMMs as well as activity group HMMs. For testing an activity in a depth video, the spatiotemporal features from the testing video are applied on the trained group HMMs and one group is selected that represents the highest likelihood. Finally, the same features are applied on the trained activity HMMs of the recognized activity group obtained using aforementioned step and the activity HMM is chosen based on the highest likelihoods. The overall proposed HAR approach is depicted in Fig. 1.

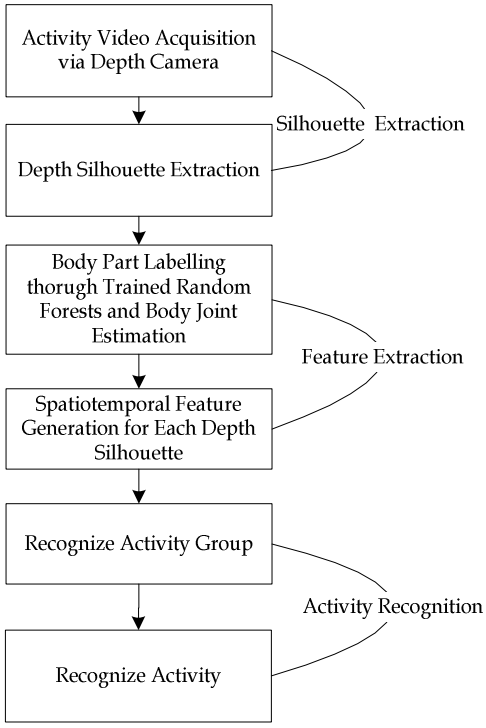


Figure 1. *Steps of the proposed system.*

II. BODY PART SEGMENTATIONS

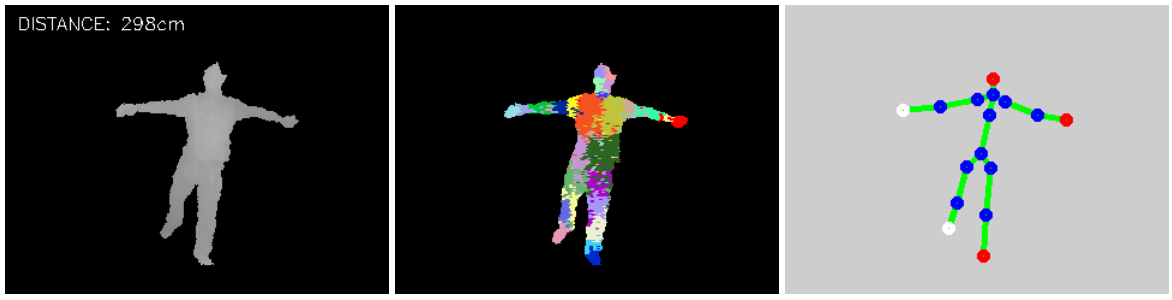
Kinect, a commercial camera is utilized here to get the depth images [9] and the depth silhouette is extracted from every depth image. Random Forests (RFs) is a useful tool to deal with multiple classification problems [10]. A forest consists of decision trees where each tree leaves represents the probability of the body labels using corresponding tree. For training RFs, simple features are used for each pixel in a depth silhouette image based on differences between different pixel pairs in window considering the pixel as center. Then, all features of all depth pixels and corresponding labels obtained from training depth silhouette images are used to train RFs. The trained RFs are used to label each pixel in the testing activity depth silhouettes. After labeling 31 body parts, a skeleton body model is generated consisting of 16 body joints. Fig. 3 shows a sample activity body silhouette, labeled

body parts, and corresponding joints from both hand waving, right hand waving, left leg up-down, and right leg up-down activity

(a)

(b)

(c)



(d)

Figure 2. *A sample activity body silhouette, labeled body parts, and corresponding joints from (a) both hand waving, (b) right hand waving, (c) left leg up-down, and (d) right leg up-down activity.*

III. HUMAN ACTIVITY FEATURE GENERATION

As aforementioned, once the labeled silhouette is available, a skeleton model representing 16 joint positions is obtained. The body parts and 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. These 16 joints are used later to obtain the features to be used with HMMs. Fig. 4 shows a sample activity depth silhouette, corresponding segmented body parts with different colors, and joints from both hand waving, right hand waving, sitting activity respectively.

The first feature information is spatial feature represented by the body joint location. Hence, considering each joint as (J_x, J_y, J_z), the spatial feature size for each image becomes 1x48. Furthermore, temporal features representing motion parameters i.e., magnitude, translation, and directional angles of the joints in the next frame are computed for 16 joints. The magnitude G of a joint from two consecutive depth frames is as

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

The translation features T and angle features E of body joint between two consecutive frames are computed as

$$T_x = J_{x(i-1)} - J_{x(i)} \quad (2)$$

$$T_y = J_{y(i-1)} - J_{y(i)} \quad (3)$$

$$T_z = J_{z(i-1)} - J_{z(i)} \quad (4)$$

$$E_{J(x,y)} = \arctan\left(\frac{J_{y(i-1)} - J_{y(i)}}{J_{x(i-1)} - J_{x(i)}}\right), \quad (5)$$

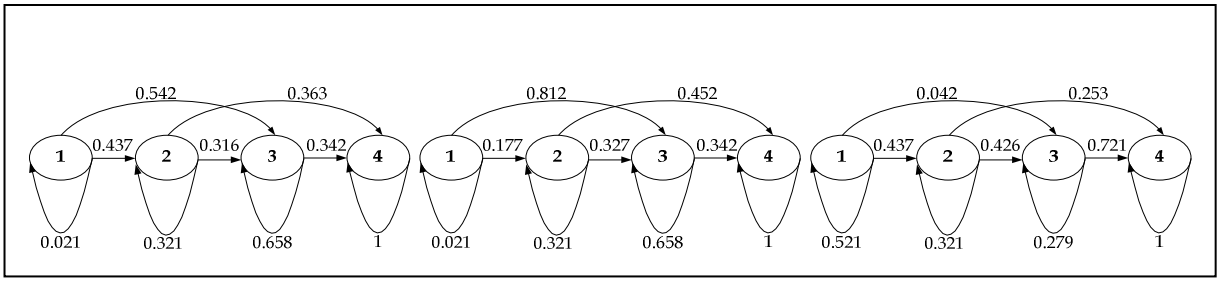
$$E_{J(y,z)} = \arctan\left(\frac{J_{z(i-1)} - J_{z(i)}}{J_{y(i-1)} - J_{y(i)}}\right), \quad (6)$$

$$E_{J(x.z)} = \arctan\left(\frac{J_{z(i-1)} - J_{z(i)}}{J_{x(i-1)} - J_{x(i)}}\right). \quad (7)$$

The size of the directional angles of the joints' becomes a vector of 1×48. Hence, the spatiotemporal features for each depth shape becomes with the size 1x112 altogether. The feature vector for i^{th} video frame is represented as F_i.

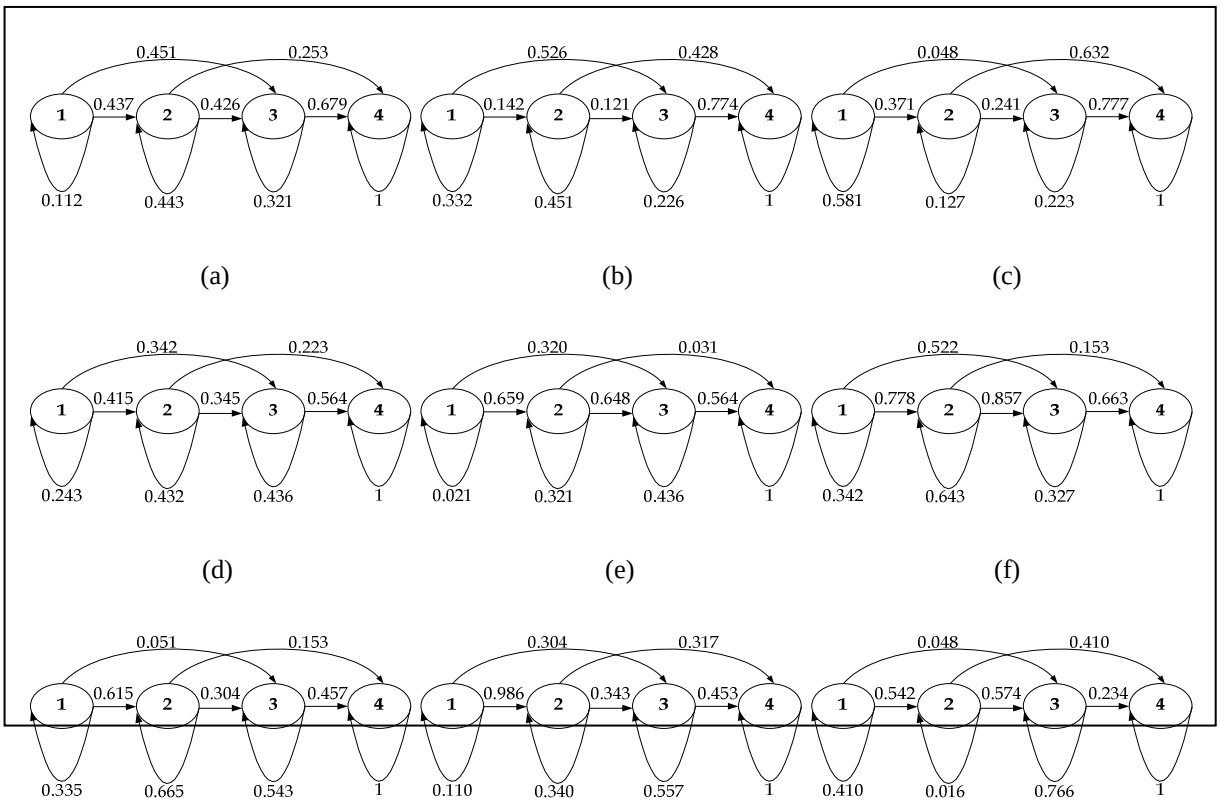
IV. HUMAN ACTIVITY MODELING AND RECOGNITION

HMM is a collection of finite states where each state has transition probability to other ones as well as observation probability of all the symbols. Recently, HMM has been attracted a lot by many researchers to decode time sequential events [13]-[15]. Two levels of activity training and testing is done. From an activity database, *K* different activity group HMMs are trained first. Then, individual activity in each group is trained. For instance, there are *N* activities in each group and hence *N* activity HMMs *H* are trained for each group. Further information regarding basic HMM is available [1], [13]-[16]. Fig. 3 shows trained HMMs of three different activity groups (i.e., first level HMMs) and Fig. 4 depicts nine trained activity HMMs (i.e., second level HMMs) of all groups used in this work.



(a) (b) (c)

Figure 3. First level trained HMMs of (a) Group1, (b) Group2, and (c) Group3.



(d) (e) (f)

(g) (e) (h)

Figure 4. Second level trained HMMs of (a) right hand up-down, (b) waving right hand to indicate to come, (c) waving right hand to indicate bye, (d) left hand up-down, (e) waving left hand to indicate to come, (f) left leg up-down, (g) both hands up-down, (h) waving both hands to indicate to come, and (i) right leg up-down.

For testing an activity video, the symbol sequence S obtained from the activity depth image sequence is applied on K trained activity group HMMs (i.e., first level HMMs) and decision U is obtained as

$$U = \arg\max_{i=1}^{K}(P(S \mid A_k)). \quad (8)$$

Then, the same sequence S is applied on the activity HMMs (i.e., some specific trained activity HMMs from second level HMMs in the activity group U) of the group U obtained in aforementioned step and the activity is chosen finally based on the maximum likelihood from those specific second level HMMs H to make the decision as

$$Decision = \arg\max_{i=1}^{N}(P(S \mid H_k)) \quad (9)$$

where $H_1, H_2, \ldots, H_N$, belongs to activity group U determined from (8).

V. EXPERIMENTS AND RESULTS

A database of nine different activities distributed in three different groups was built for training and testing. Group 1 consists of right hand up-down, waving right hand to indicate to come, and waving right hand to indicate bye. Group 2 consists of left hand up-down, waving left hand to indicate to come, and left leg up-down. Group3 consists of both hands up-down, waving both hands to indicate to come, and right leg up-down. Ten clips from each activity were used to build the training feature space where each clip consists of clip was of 15 length. Hence, 450 silhouettes were used for each group of activities. Hence, the whole training database contained a total of 1350 depth and segmented body shapes individually. Twenty five activity videos were used to test where group was determined first and activities in that specific group were used to obtain final activity decision.

The experiments were started with the conventional binary silhouette-based HAR where PCA and ICA with HMMs were tested. Since the binary silhouettes represent poor representations of the shapes (i.e., only black and white), the recognizer produced very poor recognition rates as shown in Tables I and II. Tables III and IV show recognition performance using PCA and ICA features on the depth silhouettes with HMMs where the maximum mean recognition rate was 86.11% using ICA-based local feature-based approach. Finally, the experiments were done with the proposed approach where much better recognition performance than the binary as well as depth silhouette-based experiments was obtained as shown in Table V (i.e., 93.06). Hence, the proposed approach shows the superior recognition performance over other approaches.

TABLE I. *EXPERIMENTAL RESULTS USING PCA-BASED BINARY SILHOUETTE FEATURES.*

Group	Activity	Recognition Rate	Mean

1	Right hand up-down	70	74.44
	Waving right hand to indicate to come	75	
	Waving right hand to indicate bye	72.50	
2	Left hand up-down	80	
	Waving left hand to indicate to come	77.50	
	Left leg up-down	70	
3	Both hands up-down	75	
	Waving both hands to indicate to come	72.50	
	Right leg up-down	77.50	

TABLE II. *EXPERIMENTAL RESULTS USING ICA-BASED BINARY SILHOUETTE FEATURES.*

Group	Activity	Recognition Rate	Mean
1	Right hand up-down	77.50	79.44
	Waving right hand to indicate to come	80	
	Waving right hand to indicate bye	75	
2	Left hand up-down	82.50	
	Waving left hand to indicate to come	80	
	Left leg up-down	80	
3	Both hands up-down	82.50	
	Waving both hands to indicate to come	77.50	
	Right leg up-down	80	

TABLE III. *EXPERIMENTAL RESULTS USING PCA-BASED DEPTH SILHOUETTE FEATURES.*

Group	Activity	Recognition Rate	Mean
1	Right hand up-down	75	78.06
	Waving right hand to indicate to come	77.50	
	Waving right hand to indicate bye	75	
2	Left hand up-down	80	
	Waving left hand to indicate to come	80	
	Left leg up-down	77.50	
3	Both hands up-down	80	
	Waving both hands to indicate to come	77.50	
	Right leg up-down	80	

TABLE IV. EXPERIMENTAL RESULTS USING ICA-BASED DEPTH SILHOUETTE FEATURES.

Group	Activity	Recognition Rate	Mean
	Right hand up-down	87.50	

1	Waving right hand to indicate to come	85	86.11
	Waving right hand to indicate bye	87.50	
2	Left hand up-down	82.50	
	Waving left hand to indicate to come	87.50	
	Left leg up-down	85	
3	Both hands up-down	90	
	Waving both hands to indicate to come	90	
	Right leg up-down	80	

TABLE V. EXPERIMENTAL RESULTS USING PROPOSED APPROACH.

Group	Activity	Recognition Rate	Mean
1	Right hand up-down	92.50	93.06
	Waving right hand to indicate to come	90	
	Waving right hand to indicate bye	95	
2	Left hand up-down	92.50	
	Waving left hand to indicate to come	97.50	
	Left leg up-down	92.50	
3	Both hands up-down	95	
	Waving both hands to indicate to come	92.50	
	Right leg up-down	90	

VI. CONCLUSION

In this work, a two-level HMM-based novel approach has been presented for depth video-based human activity recognition that utilizes spatiotemporal features of body joints obtained through labeling of different body parts using random forests. The experimental results show significantly improved recognition performance using proposed approach over conventional approaches. The proposed approach should be applicable in various smart environments such as smart homes.

ACKNOWLEDGMENT

This paper was supported by Faculty Research Fund, Sungkyunkwan University, 2013.

REFERENCES

[1] 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.

[2] A. Jalal, M. Z. Uddin, J. J. Lee, and T.-S. Kim, "Recognition of Human Home Activities via Depth Silhouettes and R Transformation for Smart Home," Indoor and building Environment, vol. 21(1), pp. 184-190, 2012.

[3] M. Z. Uddin, J.J. Lee, and T.-S Kim, "Independent shape component-based human activity recognition via Hidden Markov Model," Journal of Applied Intelligence, pp. 193-206, 2010.

[4] 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.

[5] V. Ferrari, M.-M. Jimenez, and A. Zisserman, "2D Human Pose Estimation in TV Shows," Visual Motion Analysis, LNCS 5604, pp. 128-147, 2009.

[6] 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, pp. 1-13, 2008.

[7] J. Wright, and G. Hua, "Implicit Elastic Matching with Random Projections for Pose-Variant face recognition," IEEE conf. on Computer Vision and Pattern Recognition, pp. 1502-1509, 2009.

[8] A. Bosch, A. Zisserman, and X. Munoz, "Image classification using random forests and ferns," IEEE Int. Conf. on Computer Vision, pp. 1-8, 2007.

[9] Microsoft Corporation: 'Kinect for Xbox 360-Xbox.com', http://www.xbox.com/en-GB/kinect/, Accessed September 4, 2014.

[10] V. Lepetit, and P. Fua, "Keypoint recognition using randomized trees," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 28, pp. 1465-1479, 2006.

[11] T. Kanungu, D. M. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Y. Wu, "The analysis of a simple k-means clustering algorithm", in Proceedings of 16th ACM Symposium On Computational Geometry, pp. 101-109, 2000.

[12] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design", IEEE Transaction on Communications, vol. 28, no. 1, pp. 84-94, 1980.

[13] E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains", Annals of Mathematical Statistics, vol. 41, pp. 164-171, 1970.

[14] E. Baum and J. Eagon, "An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology", American Mathematical Society Bulletin, vol. 73, pp. 360-363, 1967.

[15] R. Lawrence and A. Rabiner, "Tutorial on hidden markov models and selected applications in speech recognition", in Proceedings of the IEEE, vol. 77, No. 2, pp. 257-286, 1989.