2017 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI 2017)
November 16 - 18, 2017, Daegu, Korea

Human Activity Recognition Using Robust Spatiotemporal Features and Convolutional Neural Network

Md. Zia Uddin, *Senior Member, IEEE,* Weria Khaksar, and Jim Torresen, *Senior Member, IEEE*

***Abstract*—In this work, we propose a novel human activity recognition method from depth videos using robust spatiotemporal features with convolutional neural network. From the depth images of activities, human body parts are segmented based on random features on a random forest. From the segmented body parts in a depth image of an activity video, spatial features are extracted such as angles of the 3-D body joint pairs, means and variances of the depth information in each part of the body. The spatial features are then augmented with the motion features such as magnitude and direction of joints in next image of the video. Finally, the spatiotemporal features are applied to a convolutional neural network for activity training and recognition. The deep learning-based activity recognition method outperforms other traditional methods.**

Keywords—Depth videos, segmentation, skeleton, CNN.

I. INTRODUCTION

VIDEO camera-based human activity recognition has a wide range of applications in human-machine interaction and image processing [1]. For human activity feature extraction, 2-D binary shapes have been very commonly applied so far [1]-[3]. A major limitation of the binary shapes is that they cannot show the clear differences between the distant and near body parts in activity images. However, depth shapes can solve this problem and hence can be used to develop a robust activity recognition system [3]. In addition to human activity analysis in videos, body part segmentation is also grabbing good attention by many researchers [4]-[8]. For example, in [4], the authors used k-means algorithm to segment different human body parts. In [5], the authors segmented the upper body parts for the estimation of different activity poses.

For various machine learning applications, depth images have been used by a lot of researchers [1], [9]-[25]. In [9], the authors extracted 3-D point-based features from depth images and applied for activity recognition. In [10], the authors used histograms of surface orientation on depth images for human activity analysis. In [11], the authors analyzed 3-D activity patterns based on random occupancies. In [12], the authors extracted temporal motion energies from human activity depth images. In [13], the authors used depth and colour images for analysis of kitchen activities. In [14], the authors applied depth image-based features with Hidden Markov Models (HMMs) to model some activities. In [15], the authors extracted eigenfeatures from colour and depth images for activity recognition. In [16], the authors analyzed hand pose from sparsely distributed information in different activity images.

To model patterns in raw data, Deep Neural Network (DNN) has gained remarkable attention from many researchers these days [17], [18]. To train and recognize an event from given inputs, DNN can generate features from the raw inputs that makes it more robust than typical neural networks. However, DNN needs much more training time than typical neural networks, which is one of its major disadvantages. Furthermore, it often results in overfitting [17]. Recently, Convolutional Neural Networks (CNN) has become popular due to its improved discriminative power compared to other deep learning methods such as Deep Belief Network (DBN). Basically, CNN is a kind of deep learning consisting of feature extractions as well as some convolutional stacks to create a progressive hierarchy of abstract features. The different necessary parts of a CNN include convolution, pooling, tangent squashing, rectifier, and normalization [18]. As CNN seems to be a good candidate for activity pattern analysis, it is adopted in this work for robust activity recognition.

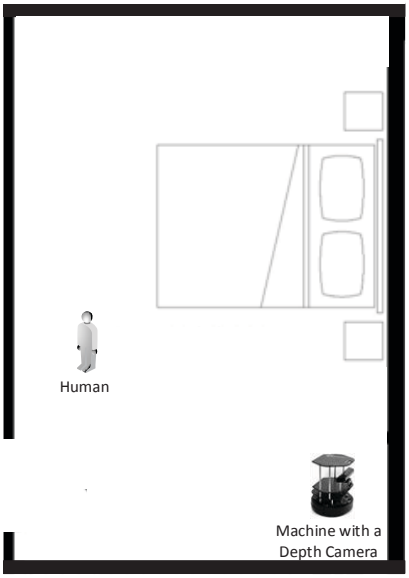


Fig. 1. A schematic room setup for depth camera-based activity recognition.

II. METHODOLOGY

The proposed human activity recognition system consists of several steps including depth video processing, feature

The authors are with the Dept. of Informatics, University of Oslo, Oslo, Norway. Emails: {mdzu, weriak, jimtoer}@ifi.uio.no.

This work is partially supported by The Research Council of Norway as a part of the Multimodal Elderly Care Systems (MECS) project, under grant agreement 247697.

generation, and modeling activity via CNN. Fig. 1 shows a schematic setup of a smart room where a mobile robot observes human activities. Fig. 2 represents the basic flows of training and testing procedures in the system. Fig. 3 shows a scene without a subject, with a subject, the overall depth image, and the subject's depth shape, respectively.

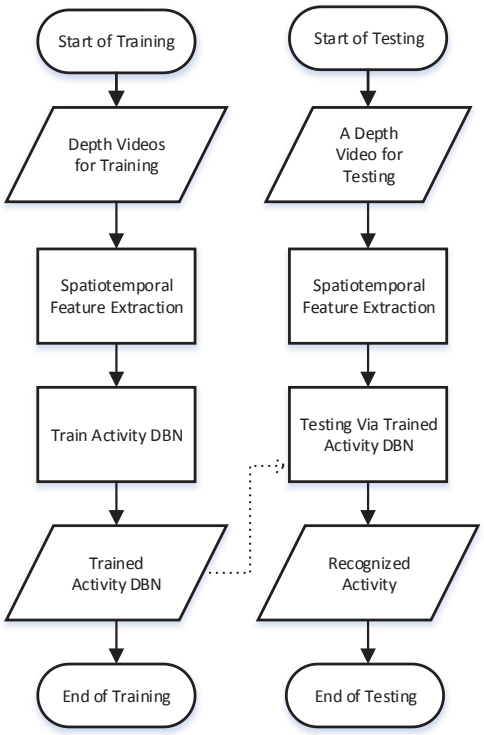


Fig. 2. Basic flows of the proposed activity recognition system.

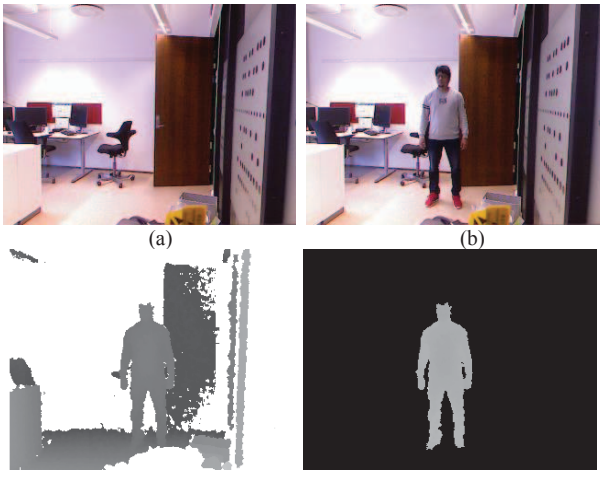


Fig. 3. (a)An example scene without a subject, (b) with a subject, (c) corresponding depth image of (b), and (d) depth shape of the subject extracted from (c).

A. Body Part Segmentation

A random forest used in this work for the segmentation process of different body parts. The forst consists of some decision trees. Each tree in the forest consists of nodes and leaves. Fig. 4 shows an example of a random forest with three decision trees. Thus, features for a decision from a depth pixel *m* can be obtained as

$$S(\alpha) = \left[D\left(\alpha + \frac{k}{D(\alpha)} \right) - D\left(\alpha + \frac{l}{D(\alpha)} \right) \right] \quad (1)$$

where $D(\alpha)$ is the depth value at pixel α, k represents a random offset value for row and l a random offset value for column around the pixel. Thus, random features and labels are obtained from all training depth pixels and used to train the random forest. The trained forest is applied to assign a label to each pixel. A synthetic database of segmented body parts and depth shapes is built to train a random forest. Fig. 5 shows a segmented body shape and corresponding skeleton. Each shape consists of 22 body parts. Fifty random features are considered at each node of a decision tree in the forest. The probabilities of the 22 body parts are considered at the leaf nodes of each tree. Finally, voting of all the trees is considered to assign a body part label to a depth pixel. At last, a skeleton model with 16 joints is obtained by considering the center of the segmented body parts.

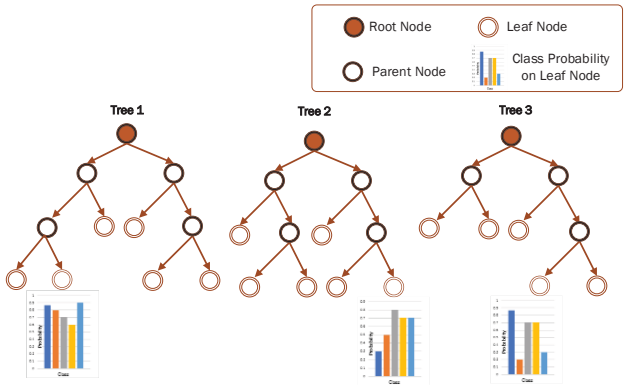
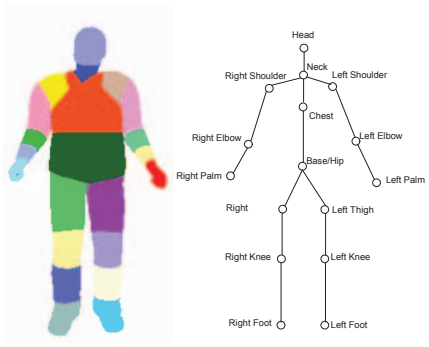


Fig. 4. A random forest structure used to train the labels of the depth silhouettes.



(a) (b)

Fig. 5. (a)A segmented body silhouette and (b) skeleton consisting of joints.

B. Feature Generation

Once we obtain the segmented body shape, the 3-D centroid of each body part is combined to represent the 3-D skeleton model. The 3-D location of each joint is denoted

by Q_{Jx}, Q_{Jy}, Q_{Jz} where J is the joint with 3-D coordinates x, y, and z. The body joints considered in this work, are the head, neck, chest, left shoulder, right shoulder, left elbow, right elbow, the centre of the hip, left hip, right hip, left palm, right palm, left knee, right knee, left foot, and right foot. After initial segmentation, a threshold is applied to the area of the labeled body parts. If a labeled body part is found in more than one places after segmentation, same labeled areas situated at very close distances are merged. For other same-labeled areas that are not close to each other and in the middle of other big labels, they are labeled with those big labels. Hence, the body part segmentation process is tried to be corrected as much as possible to get proper body skeleton. Fig. 6 shows sample body shapes with labeled body parts and skeletons overlaid on corresponding depth shapes from both hand waving, right hand waving, and sitting activities.

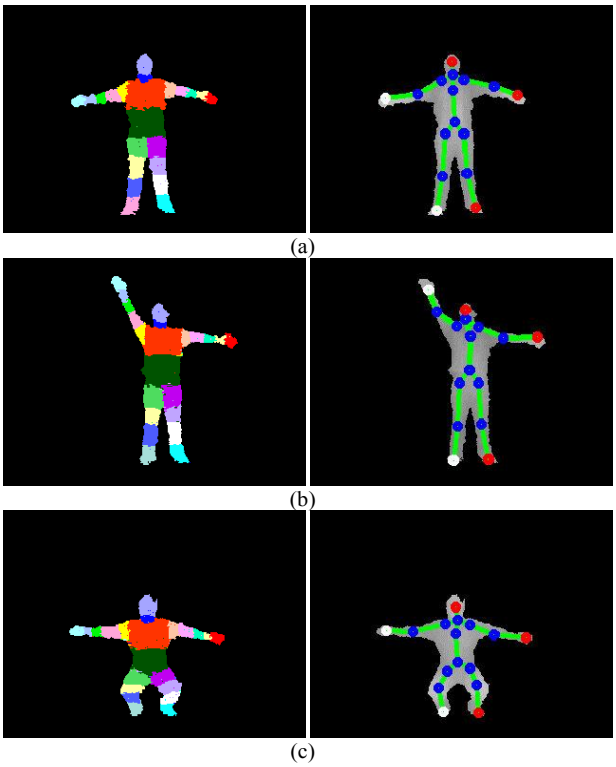


Fig. 6. A sample activity labeled body parts, and the corresponding skeleton overlaid on the depth image from (a) both hands waving, (b) right-hand waving, and (c) sitting activity.

2.2.1 Spatial Features

From a depth image, the first spatial features are represented by the angles from one joint to another. As there are 16 joints in a skeleton, there are 15*15=225 possible joint pairs. The spherical coordinate system is used to represent a joint pair with polar and azimuthal angles to make the skeleton size-invariant. The polar and azimuthal angles for a pair consisting of Q_J and Q_K joints can be obtained as follows:

$$polar = \arctan\left(\frac{Q_{Ky} - Q_{Jy}}{Q_{Kx} - Q_{Jx}}\right), \tag{2}$$

$$Azimuthal = \arcsin\left(\frac{Q_{Kz} - Q_{Jz}}{M}\right), \tag{3}$$

$$M = \sqrt{(Q_{Kx} - Q_{Jx})^2 + (Q_{Ky} - Q_{Jy})^2 + (Q_{Kz} - Q_{Jz})^2}. \tag{4}$$

Thus, all 225 pairs are considered in the spatial feature extraction process. The next feature is the mean of the depth values $\bar{X}_S$ from each body part S such that

$$\bar{D}_S = \frac{1}{W_S}\sum_{i=1}^{W_S} D_S(i) \tag{5}$$

where W_S is the number of pixels in the S^{th} body part and D_S the depth values of that part. The next feature is the variance C_S of depth values of a body part as

$$C_S = \frac{1}{W_S}\sum_{i=1}^{W_S} (D_S(i) - \bar{D}_S)^2. \tag{6}$$

The last spatial feature is calculated by adding the depth values of each body part as

$$E_S = \sum_{i=1}^{W_S} D_S(i). \tag{7}$$

Hence, the size of the spatial features for a depth shape becomes 1x538: considering 225 joint-pair angles (i.e., 1x450), the mean of the depth pixel intensities of the 22 body parts (i.e., 1x22), the variances of the depth values of the 22 body parts (i.e., 1x22), and the areas of the depth values of the 22 body parts (i.e., 1x22). Then, the spatial features are augmented horizontally for a frame. Thus, the spatial features for t^{th} frame in a video can be represented as

$$S_t = [\underset{\substack{k=1\\j=1}}{\overset{15}{\underset{15}{Polar_{jk}, Azimuthal_{jk}}}}], [\underset{s=1}{\overset{22}{\bar{D}_s}}], [\underset{s=1}{\overset{22}{\bar{C}_s}}], [\underset{s=1}{\overset{22}{\bar{E}_s}}]. \tag{8}$$

2.2.2 Temporal Features

For an activity frame, the temporal features are extracted considering the magnitude and direction of the 16 body joints in next frame. The magnitude R of a joint J at t^{th} frame is represented by

$$R_t^J = \sqrt{(Q_{Jx(t+1)} - Q_{Jx(t)})^2 + (Q_{Jy(t+1)} - Q_{Jy(t)})^2 + (Q_{Jz(t+1)} - Q_{Jz(t)})^2}. \tag{9}$$

Thus, the size of the magnitude features for each depth image becomes a vector as 1×16. The angles of a body joint J for t^{th} frame are computed as

$$A_{Q_{J(x,y)}} = \arctan\left(\frac{Q_{Jy(t+1)} - Q_{Jy(t)}}{Q_{Jx(t+1)} - Q_{Jx(t)}}\right), \tag{10}$$

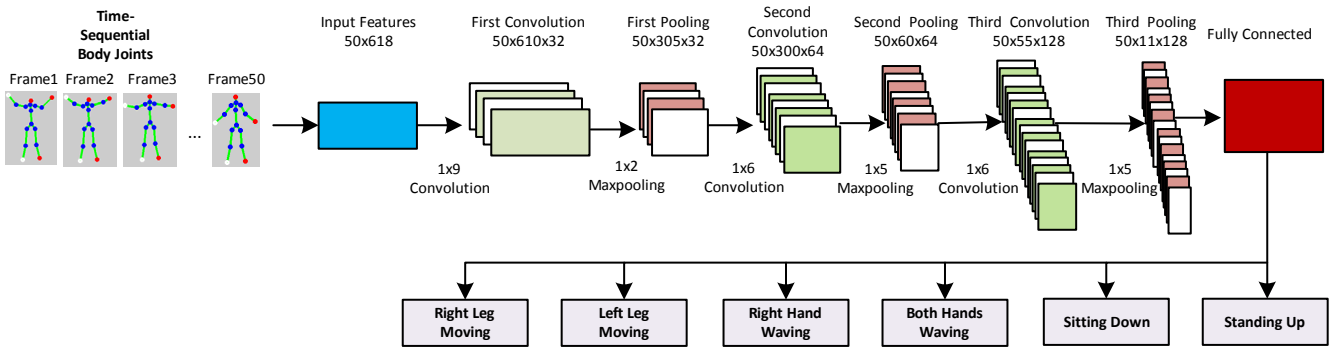


Fig. 7. Basic architecture of a 1-D CNN used in this work.

$$A_{Q_{J(y,z)}} = \arctan\left(\frac{Q_{Jz(t+1)} - Q_{Jz(t)}}{Q_{Jy(t+1)} - Q_{Jy(t)}}\right), \quad (11)$$

$$A_{Q_{J(z,x)}} = \arctan\left(\frac{Q_{Jx(t+1)} - Q_{Jx(t)}}{Q_{Jz(t+1)} - Q_{Jz(t)}}\right). \quad (12)$$

The size of the directional features for each depth image becomes a vector as 1×48. Then, the average motion of all the body joints for t^{th} frame in a video is considered as

$$\bar{R}_t = \frac{1}{N}\sum_{j=1}^{N} R(j) \quad (13)$$

where N is the number of joints. In our case N=16. Hence, the size of the temporal features for each frame is 1x80. Thus, the spatial features for t^{th} frame in a video can be represented as

$$T_t = R_t^j, [\overset{15}{\underset{j=1}{A_{Q_{J(x,y)}}}}, A_{Q_{J(y,z)}}, A_{Q_{J(z,x)}}], \bar{R}_t. \quad (14)$$

where all the features are augmented horizontally.

2.2.2 Spatiotemporal Features

For a frame in a depth video, the spatial and temporal features extracted from the body silhouette are augmented horizontally as

$$Y_t = S_t, T_t. \quad (15)$$

The augmentation makes the size of the spatiotemporal features for a frame as 1x618. For a depth video, the spatiotemporal features from each frame in the video are concatenated vertically to represent the features as

$$V = \begin{bmatrix} Y_1 \\ Y_2 \\ , \\ , \\ Y_N \end{bmatrix} \quad (16)$$

where N is the length of the video. For a video of fifty frames, the size of the spatiotemporal features becomes 50x618.

C. Convolutional Neural Network for Activity Modeling

Convolutional Neural Network (CNN) is mostly used for image-based deep learning applications [27]. Compared to other deep learning structures, CNN often demonstrates better recognition performance in machine vision applications due to its ability to extract and learn image-based features. Besides, CNN has the advantage of using a small amount of bias and weight values than other deep learning methods.

Fig. 7 depicts the structure of the 1-D CNN adopted in this work. Features from an activity video of 50 frames are organized as 50x618 and used as inputs to the CNN. The CNN has three convolution layers, three pooling layers, and one fully connected layer. The output of the network is classified into six activities shown in the figure. In first convolution layer, the input matrix is convolved with 32 convolution kernels where the size of each kernel is 1x9. As a result, a matrix of 50x610x32 is generated. The second layer is the first pooling layer where it down samples its input to a matrix of 50x305x32 via 1x2 max pooling. Using similar way, second convolution layer applies 64 convolution kernels with the size of 1x6. Third convolution layer applies 128 kernels with the size of 1x6. In second and third pooling layers, 1x5 max-pooling is used. Thus, the output of third pooling is a matrix of 50x11x128. Finally, this matrix is used as input to the fully connected layer to take a final decision. To run the CNN, the weight and bias values are initialized with random numbers. The convolution layer can be represented as

$$Convolution_i^{(j+1)}(x,y) = f(\alpha), \quad (17)$$

$$f(\alpha) = \sum_{m=1}^{n} \Omega(x,(y - m + \frac{n+1}{2}))\Xi_i^j(m) + \Theta_i^i \quad (18)$$

where $Convolution_i^{(j+1)}$ represents the convolution results to $(j+1)^{th}$ layer with i^{th} convolution map. Ξ_i^j is i^{th} kernel for j^{th} layer. Θ_i^j represents i^{th} bias values for layer j. n is the

size of the kernel and Ω the map of the previous layer. The output of pooling layer *j* at coordinate *(m,n)* can be represented as

$$Maxpool^{j}(m,n) = \max_{g=1,2,..,w} (\Omega(m,((n-1)*w+g)). \quad (19)$$

where *w* is the length of the pooling window. Finally, the fully connected layer can be represented as

$$Fully\_Connected_{j}^{(l+1)} = f(\sum_{i} z_i^l \Xi_{ij}^{\ l} + \Theta_j^l) \quad (20)$$

where $\Xi_{ij}^{\ l}$ is the weight matrix from the i^{th} node of the l^{th} layer to the j^{th} node of the $(l+1)^{th}$ layer. z_i^l is the content of i^{th} node at l^{th} layer. Thus, 1-D CNN is applied on the robust spatiotemporal features for human activity recognition.

III. Experiments and Results

A database of six activities was built to check different human activity recognition approaches. The activities were left leg moving, right leg moving, both hand waving, right hand waving, sitting-down, and standing-up. One hundred clips from each activity were collected for the training purpose. Finally, 100 clips were used to test each activity.

TABLE I. Experimental results on binary silhouettes using different approaches.

Approach	Activity	Recognition Rate (%)	Mean (%)
PCA on Binary Silhouette-Based Activity Recognition	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 on Binary Silhouette-Based Activity Recognition	Right Leg Moving	79	**77.33**
	Left Leg Moving	74	
	Right Hand Waving	81	
	Both Hand Waving	73	
	Sitting-Down	79	
	Standing-Up	83	

We started the experiments with the conventional binary and depth shape-based activity recognition approaches with HMM. As binary silhouettes represent only binary colour (i.e., black and white) representations, the recognizer showed very poor recognition performance, as shown in Table I where the highest mean recognition rate was 77.33% using independent component analysis (ICA) [3]. Basically, ICA represents better features than principal component analysis (PCA). The experiments were then continued to the depth silhouette-based activity recognition. Table II shows the experimental results where it indicates the superiority of the depth silhouettes over the binary ones.

TABLE II. Experimental results on depth silhouettes using different approaches.

Approach	Activity	Recognition Rate (%)	Mean (%)
PCA on Depth Silhouette-Based Activity Recognition	Right Leg Moving	75	**76.67**
	Left Leg Moving	77	
	Right Hand Waving	83	
	Both Hand Waving	81	
	Sitting-Down	71	
	Standing-Up	73	
ICA on Depth Silhouette-Based Activity Recognition	Right Leg Moving	87	**87.33**
	Left Leg Moving	85	
	Right Hand Waving	91	
	Both Hand Waving	89	
	Sitting-Down	83	
	Standing-Up	89	

TABLE III. Experimental results using Spatiotemporal features using different approaches.

Approach	Activity	Recognition Rate (%)	Mean (%)
Spatiotemporal Feature-based Activity Recognition 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	
Proposed Spatiotemporal Feature -based Activity Recognition with CNN	Right Leg Moving	99	**98.17**
	Left Leg Moving	97	
	Right Hand Waving	99	
	Both Hand Waving	99	
	Sitting-Down	97	
	Standing-Up	98	

Finally, the spatiotemporal feature-based experiments were done where significantly improved recognition performances were obtained, as included in Table III. Firstly, the spatiotemporal features were combined with HMM which achieved 91.33% mean recognition rate. Later on, the proposed approach (i.e., the spatiotemporal features with CNN) was tried that achieved the highest recognition rate (i.e., 98.17%), showing its superiority over all other approaches.

A. Experiments on MSRC-12 Gesture Dataset

The proposed approach was checked on MSRC-12 gesture dataset too [28]. The dataset consists of sequences of human skeletal joint movements. It has 594 sequences collected from 30 people for twelve different activities. We created 6000 sequences using cross folding where 5000 were used for training and 1000 for testing. The activities are Lift arms, Duck, Push right, Goggles, Wind it up, Shoot, Bow, Throw, Had enough, Change weapon, Beat both, and Kick. We compared our approach with the traditional HMM-based one where the proposed one showed much better accuracy (i.e., 98.27%) than the traditional one (92.49%), as represented in

Tables IV and V.

TABLE IV. EXPERIMENTAL RESULTS USING TRADITIONAL HMM-BASED APPROACH ON MSRC-12 DATASET

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

TABLE V. EXPERIMENTAL RESULTS USING PROPOSED CNN-BASED APPROACH ON MSRC-12 DATASET

Activity	Recognition Rate (%)	Mean (%)
Lift arms	98.1	
Duck	98.8	
Push right	98.8	
Goggles	97.9	
Wind it up	98.7	
Shoot	100	**98.27**
Bow	98.8	
Throw	96.9	
Had enough	98.0	
Change weapon	96.1	
Beat both	98.7	
Kick	98.4	

IV. CONCLUSION

In this paper, a novel approach has been proposed for depth camera-based human activity recognition utilizing robust spatiotemporal features and 1-D CNN. The experimental results also showed significantly improved performance by the proposed method than the traditional methods. The proposed system can be adopted in many practical applications such as smart home elderly care systems for improving the quality of their independent life.

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] M. Z. Uddin, "Human Activity Recognition Using Segmented Body Part and Body Joint Features with Hidden Markov Models ", *Multimedia Tools and Applications*, doi:10.1007/s11042-016-3742-2, 2016.

[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*, vol. 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*, vol. 2008, 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] W Li., Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," *in Proceedings of Workshop on Human Activity Understanding from 3D Data*, pp. 9-14, 2010

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

[11] J. Wang., Z. Liu, J. Chorowski, Z. Chen, and Y. Wu, " Robust 3d action recognition with random occupancy patterns," *in Proceedings of European Conference on Computer Vision*, pp. 872-885, 2012.

[12] X. Yang, C. Zhang, and Y Tian, " Recognizing actions using depth motion mapsbased histograms of oriented gradients," *in Proceedings of ACM International Conference on Multimedia*, pp. 1057-1060, 2012.

[13] J. Lei, X. Ren, qnd D. Fox, "Fine-grained kitchen activity recognition using rgb-d," *in Proceedings of ACM Conference on Ubiquitous Computing*, pp.208-211, 2012.

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

[15] X. Yang and Y. Tian, "Eigenjoints-based action recognition using naive-bayesnearest-neighbor," *in Proceedings of Workshop on Human Activity Understanding from 3D Data,* pp. 14-19, 2012.

[16] H. Hamer, J. Gall, T. Weise, and L. Van Gool, "An object-dependent hand pose prior from sparse training data," *in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. pp. 671-678, 2010.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *NIPS*, pp. 1097-1105, 2012.

[18] F. Deboeverie, S. Roegiers, G. Allebosch, P. Veelaert and W. Philips, "Human gesture classification by brute-force machine learning for exergaming in physiotherapy," *in Proceedings of IEEE Conference on Computational Intelligence and Games (CIG),* Santorini, pp. 1-7, 2016.