A Depth Video-based Human Detection and Activity Recognition using Multi-features and Embedded Hidden Markov Models for Health Care Monitoring Systems

Ahmad Jalal[1] , Shaharyar Kamal[2] and Daijin Kim[1]

[1]Department of Computer Science and Engineering, POSTECH, Pohang, 790-784, Korea
2Department of Electronics and Radio Engineering, KyungHee, Suwon, 446-701, Korea

***Abstract* — Increase in number of elderly people who are living independently needs especial care in the form of healthcare monitoring systems. Recent advancements in depth video technologies have made human activity recognition (HAR) realizable for elderly healthcare applications. In this paper, a depth video-based novel method for HAR is presented using robust multi-features and embedded Hidden Markov Models (HMMs) to recognize daily life activities of elderly people living alone in indoor environment such as smart homes. In the proposed HAR framework, initially, depth maps are analyzed by temporal motion identification method to segment human silhouettes from noisy background and compute depth silhouette area for each activity to track human movements in a scene. Several representative features, including invariant, multi-view differentiation and spatiotemporal body joints features were fused together to explore gradient orientation change, intensity differentiation, temporal variation and local motion of specific body parts. Then, these features are processed by the dynamics of their respective class and learned, modeled, trained and recognized with specific embedded HMM having active feature values. Furthermore, we construct a new online human activity dataset by a depth sensor to evaluate the proposed features. Our experiments on three depth datasets demonstrated that the proposed multi-features are efficient and robust over the state of the art features for human action and activity recognition.**

***Keywords* —Depth camera, Embedded Hidden Markov Models, Feature Extraction, Human Activity Recognition.**

I. Introduction

MONITORING human activities of daily living is an essential way of describing the functional and health status of a human [1]. Therefore, human activity recognition (HAR) is one of genuine components in personalized life-care and healthcare systems, especially for the elderly and disabled [2]. To monitor daily activities of the elderly people, video-cameras can be deployed in smart environments, such as smart homes or smart hospitals to acquire time-series activity video clips. According to the world health organization survey, the population of older people is rapidly increasing all over the world and their healthcare needs become more complex which consume more resources (i.e., human and healthcare expenditures). Thus, healthcare monitoring services are needed to overcome the extensive resource utilization and improve the quality of life of elder people [3]. Indeed, several studies support that personalized life-care and healthcare services can decrease the mortality rate especially for the elderly people. For instance, in the European countries, it is estimated that the survival proportion of older people is increasing while receiving the personalized healthcare services instead of receiving institutional care or nursing homecare [4]. Thus, the aim of this study is to propose an efficient depth video-based HAR system that monitors the activities of elder people 24 hours/day and provides them an intelligent living space which comfort their life at home.

To improve the personalized healthcare services of elderly people, automated HAR systems are required to monitor the elderly daily activities and provide safe and independent life at home [5], [6]. In automated HAR system, various sensors are used to extract signal or video data, provide continuous health monitoring by observing their daily routine activities and generate an alert in case of emergency to authorized person (i.e., doctor, nurse and relatives). Many researchers in the field of automated HAR frequently use wearable sensors or vision-based sensors to extract features data. In wearable sensor HAR technology, subjects are asked to wear sensing devices (i.e., accelerometer, magnetometer and gyroscope) at different locations of human body to capture sequence of data [7], [8]. However, HAR system built by wearable sensors have certain difficulties faced by elderly people to perform daily activities such as discomfort to wear sensors to their body parts for long time, elderly people often forget to wear proper suit equipment's of wearable sensors, directional control issues causes unreliable data recording during complex subject's movements (i.e., especially in smart phones and wrist bands/watches) and there is a relatively difficulty in terms of energy consumptions and device settings. On the other hand, video sensors have certain issues such as privacy, fixed devices at predetermined positions and pre-processing complexity. From the above literature, we have observed that video sensors have mostly solvable issues, richer information and have wider scope, therefore, we motivated to use vision-based HAR approach to recognize activities of elderly people at home.

Vision-based HAR system is a challenging research topic in the field of computer vision and pattern recognition. It enables the development of various practical applications such as security systems, elderly healthcare systems, video surveillance and smart homes systems [9], [10] which provide personal security, cost-effectiveness, friendly services and efficient health care for elderly people. In recent years, most of the researches of vision-based HAR are focused on daily activity monitoring due to the fact that less medical aid, not due attention by their relatives and loss of independence causes lives risks and injuries in elderly people. This paper annotated a novel set of continuous online daily routine activities consisting of sitting down, eating, falling-down, exercising, and taking-medicine and reading an article. To get realistic and natural scenes faced by elderly people in their life, activities are defined and selected after visiting healthcare medical centers, dealing with doctors/nurses and reviewing medical research papers [11]-[13].

DOI: 10.9781/ijimai.2017.447

II. Related Work

In this section, we will review related works from two different aspects including elderly healthcare applications and automatic activity recognition over multiple datasets.

A. Elderly Healthcare Applications

Elderly healthcare monitoring has a direct link with HAR systems, where the sensor devices capture and examine the indoor behaviors and activities of elderly at hospitals, home or offices. In smart hospitals [14], automatically estimating hospital-staff-patients activities are examined and empowered HAR to boost our vision for the hospital as a smart environment. In smart homes [15], home activities of elderly people are recognized based on invariant features characteristics. While, in smart offices [16], authors proposed comfort management system along with activity recognition solutions that handles multiple-user and rapidly recognize office activities.

B. Automatic Activity Recognition

Many existing studies have applied HAR utilizing video sensors technologies (i.e., RGB cameras) for human detection, tracking and activity recognition. In [17], a new model is proposed for activity recognition that combines a powerful mid-level representation, in the form of HoG and BoW poselets, with discriminative key frame selection based on conventional videos. In [18], an epitomic representation for modeling is introduced where the video activity sequence is divided into segments to extract moving objects and short-time motion trajectories. This information is further processed by Iwasawa matrix decomposition to represent the effect of rotation, scaling and projective action on the state vector and used for activity recognition. In [19], a view-specific approach is proposed for representation of movements as temporal templates. These templates indicate the presence of motion in binary values and the function of the motion in a sequence. Then, a matching algorithm is used to construct a recognition system. However, these cameras have certain limitations such as technical infeasibility to differentiate between near and far parts of human body, limited information (binary or RGB intensity values), highly sensitive with lighting conditions and unreliable for postures having self-occlusions.

To improve HAR capabilities and human silhouettes representation, depth sensors [20]-[22] have been released to facilitate the human detection, feature extraction and activity recognition tasks. Compared with the digital RGB cameras, depth cameras provide additional human body parts information, insensitivity to light changes and easily normalized during body orientation/size changes. In addition [21]-[26], several research articles have used depth maps information to explore their features extraction from two basic types such as depth silhouettes features and skeleton-based features for recognizing human activities using depth sequences. In the depth silhouettes features, many researches used a set of depth pixels of the depth images or human shape silhouettes to extract the features. In [20], action graph is used to model explicitly the dynamics of human motion and a bag of 3D points to characterize a set of salient postures that corresponds to the nodes in the action graph for recognizing actions/activities. In [21], a new descriptor is proposed for activity recognition using a histogram capturing the distribution of the surface normal orientation in the 4D space of time, depth, and spatial coordinates. To build the histogram, they created 4D projectors, which quantize the 4D space and represent the possible directions for the 4D normal. In [22], semi-local features called random occupancy pattern (ROP) features are proposed which employed a novel sampling scheme and extracted from randomly sampled 4D subvolumes with different sizes and locations using depth images.

Instead of relying on depth silhouettes features, many researchers have explored features based on skeleton joints information. In [23], a set of features such as body pose, hand position, motion information and point-cloud features are proposed having three dimensional Euclidean coordinates and the orientation matrix of each joint to recognition activities using RGBD images. In [24], an effective method is proposed that consists of a new type of features based on position differences of 3D joints and Eigenjoints. The Eigenjoints are able to capture the properties of posture, motion and offset of each frame. Then, they used frame descriptors of Eigenjoints without quantization and classify different actions based on Naïve Bayes Nearest Neighbor (NBNN). In [25], an actionlet ensemble model is developed to represent each action, capture the intra-class variance and recognize various actions and activities using benchmark depth datasets. Although, depth video-based HAR systems are quite feasible for recognizing activity, however, it is still difficult using just depth silhouettes features or joint point's information especially during self-occlusion. Therefore, our research work is focused on utilizing depth data based on merging both silhouettes and joint points information for feature representation, activity training and recognition.

Our main contribution of this paper is to propose a real-time body parts tracking method that has the ability to track the self-occluded human body parts especially in case of torso rotation. Also, the proposed method detects and controls the fast moving human body parts and has invariant characteristics with respect to body size and human position which strengthen our contributions in HAR. Combination of depth silhouettes and joint information features cover various factors such as robust to noises, missing joints and capture the local dependencies over the embedded HMM acting as a novel methodology in order to enhance the recognition rate over all three depth datasets. In addition, we provide a new online depth human activity dataset, which becomes a benchmark in HAR.

In this work, a novel approach is proposed for HAR by considering multi-features approach that is extracted in 3D coordinates along with time space (i.e., 3D human silhouettes and spatiotemporal joints values) and embedded HMMs. These features deal with intensity differentiation profiles, directional angular values, local motion of active body parts and temporal frame movements which provide compact and sufficient information for human action and activity recognition. All these features are concatenated together and converted into discrete symbols by considering vector quantization algorithm. Meanwhile, active regions of depth silhouettes (i.e., moving body regions, arms, legs and interest body joints) make specific classes for training and recognition using embedded HMMs. In order to determine the recognition performance, we build a new online depth activity dataset that contains segmented video sequences for training phase and unsegmented video sequences for testing phase, which will be a benchmark for activity recognition based on depth data. In addition, we evaluate our system according to the standard experimental protocols definition on three challenging depth datasets. Our results outperform all published state of the art feature extraction methods.

The rest of the paper is organized as follows. In Section III, we describe the system architecture including problem statement, dataset generation and proposed HAR system model. Section IV presents the detail description of HAR. Section V describes the experimental results and comparisons using proposed and state of the art methods. Finally, Section VI presents the conclusion of the paper.

III. System Architecture and Methodology

A. Problem Statement

Due to recognition of natural scenes of continuous human activities without any instructions to subjects, video based HAR systems faces

various problems such as dynamic backgrounds changes in a scene along with self-occlusion, object hurdles or body parts rotation (i.e., especially torso) of human subjects and body orientation or sizes changes frequently due to different subjects performing activities at different distances from cameras. Also, similar postures of different activities (i.e., falling-down and taking an object, take-medicine and eating and clapping and exercise) causes reduction in recognition performance and processing time consumption during testing of activities especially online datasets.

We proposed an online depth HAR system that utilized person-tracking system, multi-features and embedded HMMs algorithms to solve the above mentioned problems. For the first problem, we developed our real-time body tracking system to control self-occlusion and provide freely human movements in a scene. For the second problem, we normalized skeleton models and applied invariant features to manage body size changes. While, third problem is solved by considering multi-features having depth silhouettes and joints information to identify difference in between different activities having similar postures. For the fourth problem, we introduced embedded HMMs concept to overcome the redundant data usage during testing time, improve the computational processing and increase recognition performance.

B. Dataset Generation

During daily routine activities, elderly people are mainly involved in a mixture of static sequences (i.e., minor movements of body parts) and dynamic sequences (i.e., major movements of different body parts) of activities. Therefore, our dataset provided both types of activity sequences having natural routine behaviors and continuous recognition of elderly people such as eating, taking-medicine, sitting down, exercising, falling-down and cleaning. Due to monitoring of continuous elderly activities, detecting starting and ending times of all occurring activities are controlled by a sliding window approach. In Fig. 1 we annotated an online continuous depth dataset by considering the daily life activities of elderly people at home or offices.

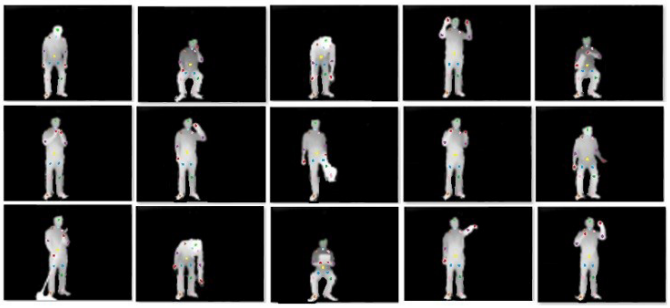

Fig. 1. Samples of human depth silhouettes along with joint point's location in our depth annotated dataset. Top row: sitting down, taking-medicine, falling-down, both hands waving, eating; Middle row: clapping, phone conversation, walking, exercising, stand up; Bottom row: cleaning, taking an object, reading an article, pointing as object and hand waving.

C. Proposed HAR System Model

The proposed framework of HAR system consists of the following processes namely as, (1) depth imaging acquisition, (2) human silhouettes segmentation and tracking, (3) feature representation and extraction based on multi-features, (4) clustering algorithm and vector quantization, and (5) human activity modeling, training and recognition. Fig. 2 shows the overall architecture of our proposed activity recognition system.

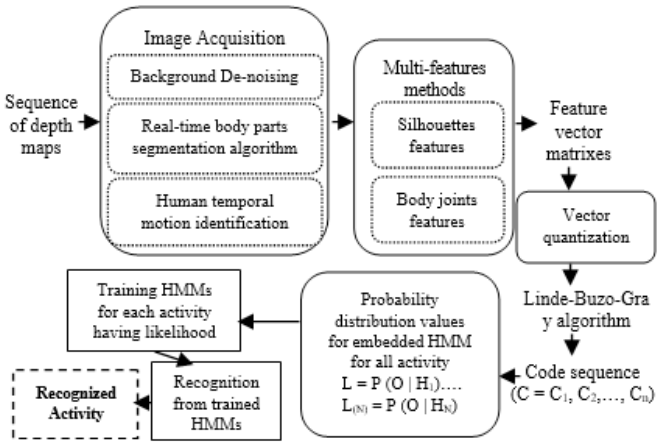

Fig. 2. System architecture of the proposed activity recognition system.

IV. Image Acquisition, Feature Extraction and Activity Recognition

In this section, we describe depth imaging acquisition, feature extraction via multi-features approach, symbol representation via Linde-Buzo-Gray (LBG) clustering algorithm and activity training/recognition using embedded HMMs.

A. Depth Imaging Acquisition

To capture 3D information, we utilized a RGB-D camera (i.e., Kinect) to acquire a pair of RGB images and depth maps. These depth maps contain a considerable amount of noise data. However, noise reduction is an important process before extracting multi-features. Therefore, to remove the noisy background areas from the depth map, we applied pixel differentiation method as

$$P_d^t(x,y,z) = b_t(x,y,z) - d_t(x,y,z) > T_{values} \quad (1)$$

Where $b_t(x, y, z)$ and $d_t(x, y, z)$ are the background and depth intensity pixel values at time t and T_{values} is a positive threshold value. Meanwhile, to apply floor removal mechanism, we simply ignore ground line (i.e., y parameters) which acts as lowest value (i.e., equal to zero) corresponding to a given pair of x and z axis. However, to extract accurate human silhouette region [26] from the scene, we calculate depth intensity center values from the scenes using connected component labeling technique (see Fig. 3). In component labeling technique, the variation of pixel intensity in an image is observed using raster scanning where every depth pixel d_n of the connected component has depth value, intensity values of two neighboring pixels are within a threshold δ_n and each object/subject has its depth center d_c values is assigned as $| d_c - d_n | \leq \delta_t$. Due to depth center values, we monitored the pixel-neighboring intensity variation in between the consecutive frames which remove the unnecessary objects (i.e., cupboard, doors or chairs) from the scenes.

Lastly, we applied human movement detection $D(f)$ by considering temporal continuity constraints between consecutive frames.

$$D(f) = \sqrt{\left(f_{t-1}^x - f_t^x\right)^2 + \left(f_{t-1}^y - f_t^y\right)^2 + \left(f_{t-1}^z - f_t^z\right)^2} \quad (2)$$

As a result, human silhouettes regions are enclosed within the rectangular bounding box having specific parametric values (i.e., height and width) based on motion detection. Fig. 3 describes the overall scenario of real time body tracking system including (a) depth maps having noisy information, (b) temporal human motion using ridge data, (c) human identification and (d) depth human silhouettes.

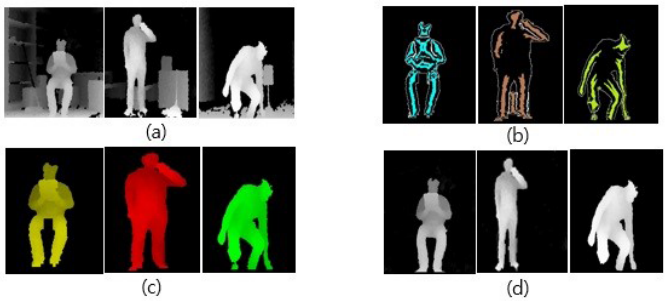

Fig. 3. Overall description of human silhouettes identification as (a) depth maps having noisy information, (b) temporal human motion using ridge data, (c) labeled human body silhouette and (d) depth human silhouettes identification of reading an article, phone conversation and falling-down activities.

B. Feature Extraction via Multi-Features Approach

In this section, we extract features from human body silhouettes and joints information (i.e., include 15 joint points) via depth images. However, a set of feature extraction techniques provide a compact representation of image content by describing invariant characteristics of local body parts, multi-view differentiation and spatiotemporal body joints motion which are derived to merge together having spatial and temporal depth silhouettes characteristics.

1) Invariant Features

To process the human depth silhouettes, we compute the total pixel intensities along lines of different locations (i.e., 0 to 180 degrees) to identify specific view directions and the center points of human silhouettes are selected as the reference point, which is defined as

$$R_D(\rho,\theta)\int_{\infty}^{-\infty}\int_{\infty}^{-\infty} f(x,y)\,\delta(\rho - x\cos\theta - y\sin\theta)\,dxdy \tag{3}$$

$R_D(\rho,\theta)$ is the line integral of the 2D radon function along a line between positive to negative infinite value. These 2D information are passed through sum of the squared Radon transform values [27] to create a 1D profiles as shown in Fig. 4. These 1D profiles are strong candidates to provide translation and scaling invariant features. Finally, a feature vector with 180 dimensions is extracted, instead of the 2D shape matrix.

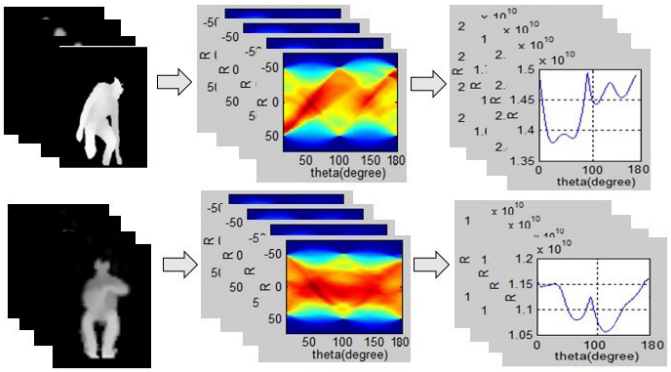

Fig. 4. Invariant features performed over various depth activities.

2) Multi-view Differentiation Features (MDF)

Due to similar postures, object occlusions and missing body parts from the frontal view, extra views (i.e., side and top) are used for the feature vectors to improve the accuracy of the classifier. Therefore, we applied Cartesian planes over the human depth silhouettes to get the 2D images of side and top views. These images are passed through a frame differential mechanism where current frame is compared with the next frame to get pixel intensity information for feature processing. Fig. 5 explains the multi-view differentiation features having (a) side and (b) top views of forward kick and bend actions using MSR Action3D dataset.

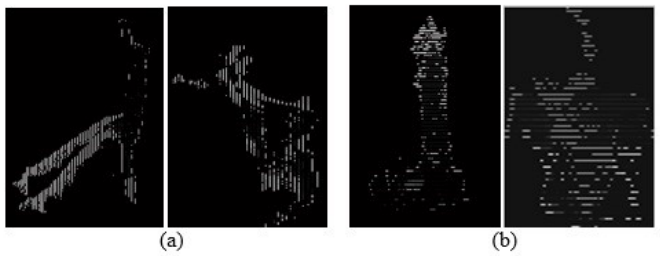

Fig. 5. Multi-view differentiation features based on (a) side and (b) top views in case of forward kick and bend actions using MSR Action3D dataset.

3) Spatiotemporal Body Parts Motion Features (BMF)

To consider the discriminative information for determining how a person has moved (i.e., spatiotemporally) during the activity sequence, we considered the shape information having specific motion region of body parts in an activity. Therefore, we calculated the gradient orientation, pixel intensity in between initial frame till final frame and Mahalanobis distance for matching the input activity from the stored templates is defined as

$$D_{seq} = \sum_{j=1}^{N} \left| I_j^t(x,y,z) - I_j^{t-1}(x,y,z) \right| > \tau \tag{4}$$

where I_j is image sequence along with t and t-1 to evaluate the temporal sequential human motion of overall activity images. However, those regions having maximally confident patches of temporal values (i.e., greater than specific threshold values) are tracked based on optical flow mechanism which are enclosed by rectangular box as shown in Fig. 6.

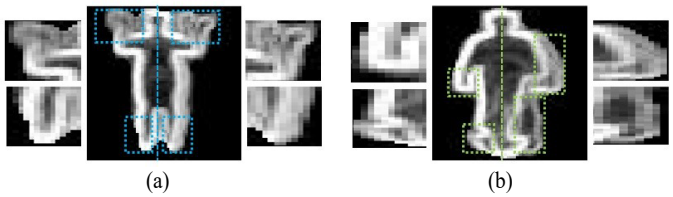

Fig. 6. Spatiotemporal body parts motion features based on maximally confident patches in case of (a) both hands waving and (b) walking activities using our depth annotated dataset.

However, both MDF and BMF features spaces produce a higher number of features dimension, thus, principal component analysis (PCA) is used here to extract global information [28] from all activities data and approximate the higher features dimension data into lower dimensional features. In this work, 200 and 150 principal components (PCs) are used for MDF and BMF features to process the activity data and are expressed $PC = m_i e_{top}$ where PC is the PCA projection of feature vectors, m_i is the zero mean vector and e_{top} is the top eigenvectors indicating higher variance among overall eigenvalues.

4) Temporal Joints Difference Features (TDF)

In addition to depth human silhouettes, the multi-features also provided skeleton motion joints features. In order to make use of the additional motion information from joints information, we applied current frame differentiation method to calculate joint points difference

between the current frame f_i and all the respective frames f_d of activity sequence can be represented as

$$f_{current}^{diff} = \left\{ f_i^j - f_r^j \mid i = 1; r = 2, \ldots, N \right\} \quad (5)$$

where 3D skeleton joints j having all three coordinates axis at frame t to $t+1,\ldots,t+N$ and the size of feature vector become 15x1, respectively.

5) Pairwise Joints Distance Features (PJF)

To consider the pairwise joint distance feature, we measure the joints distance between the active a body joints with the inner i body joints at each frame t. Here, the active body joints consist of head, shoulders, hands and feet, while, inner body joints include torso, neck, elbows, hips and knees. Thus, pairwise joints distance features D_{pjf} is represented as

$$D_{pjf}(t) = \sqrt{\left(j_a^x - j_i^x\right)^2 + \left(j_a^y - j_i^y\right)^2 + \left(j_a^z - j_i^z\right)^2} \quad (6)$$

where D_{pjf} becomes a vector of 54 dimensions. Fig. 7 shows 2D plots of TDF and PJF feature values using falling-down activity.

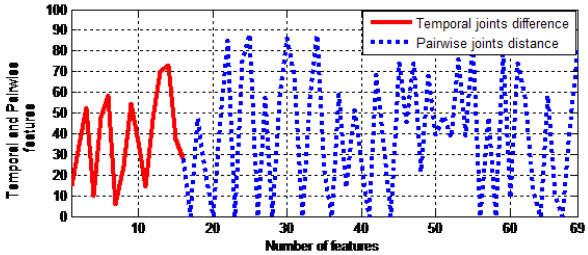

Fig. 7. Temporal joints difference and pairwise joints distance features are applied over falling-down activity.

6) Spatiotemporal Joints Angular Features

Moreover, the human body parts variation in terms of postures, direction, height and angular dimensions have a huge impact on the performance of the activity recognition. Therefore, we identify the gradients representation with respect to angles of each body joints in-between t and $t-1$ consecutive frames of each sequence can be expressed as

$$\theta_{tan} = \arctan\left(C_{(1)}^t - C_{(1)}^{t-1} / C_{(2)}^t - C_{(2)}^{t-1}\right) \quad (7)$$

$$\varphi_{cos} = arccos\left(C_k - C_k^{'} / \left|C_k\right|\left|C_k^{'}\right|\right), k = x, y, z \quad (8)$$

where $C_{(1)}$ and $C_{(2)}$ are the pair-coordinates of all three respective axis. Both equations are examined for angular and sinusoidal features characteristics. As a result, the size of the feature vector of joints angular features representation of each activity frame become 45 x 2, respectively (See Fig 8.)

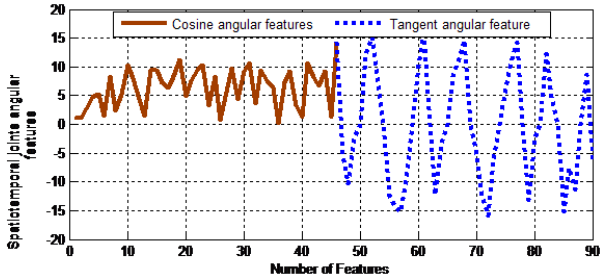

Fig. 8. 2D plot of spatiotemporal joints angular features using falling-down activity from our depth annotated dataset.

C. Symbol Representation and Code Matrix Selection

All these sub-features are merged together to make a multi-feature vector size of 689x1. These multi-features are symbolized from the activity frames and generated from Linde-Buzo-Gray (LBG) clustering algorithm. Here, LBG used a splitting mechanism where the centroid for the training activity sequence is calculated and split into two nearest vectors. Each partition is restricted into specific centroid value. Therefore, each vector is then split into two vectors and the iteration is repeated until N-level centroid values are obtained. Finally, for each cluster, samples are assigned to the same class (i.e., activity label) that is the one of the closest cluster centroid. Fig. 9 shows the internal concept of feature dimensional structure and code matrix selection of proposed features.

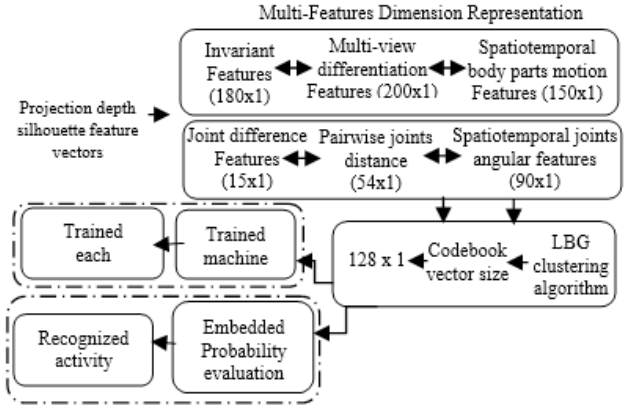

Fig. 9. Symbol representation and code matrix selection.

D. Embedded Hidden Markov Model

To model, train and recognize different activities using depth data, we introduced a new concept of embedded HMM method. Therefore, embedded HMM is introduced which focused specifically at active feature regions of human body joints such as hands, head, feet and shoulders.

Also, it includes the overall human silhouettes information or all joints information which contain redundant information such as static body regions (i.e., torso, chest and forearms) and inactive body joints (i.e., elbows, neck and hips). These kinds of unnecessary information causes reduction during performance of recognition accuracy results.

Also, full-body silhouettes contain specific or active feature regions (i.e., moving body parts areas) which are augmented together to build a single HMM having O_a observation probabilities of M active feature regions of each activity as a.

$$A_l = \sum_{l=1,2,\ldots,N,a=1}^{M} P\left(O_a | h_{la}\right) \quad (9)$$

where A_l indicates the likelihood of l HMM with respect to N number of activities. Fig. 10 shows active feature regions of overall human silhouettes to calculate specific likelihood of each activity. Finally, recognized activity R is chosen as desired activity having maximum likelihood values [29] among all activities during testing.

$$R_{act} = argmax\left\{A_l / l\right\} \quad (10)$$

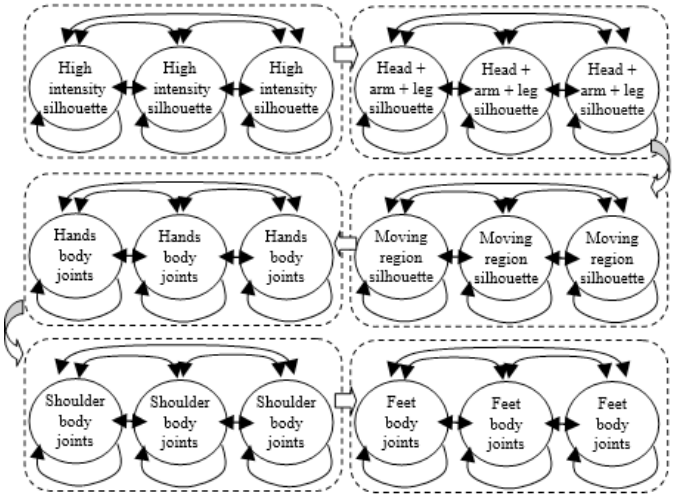

Fig. 10. Embedded HMM structure for all activities or actions using specific full-body silhouettes and joints information.

V. Experimental Results and Discussion

In this section, we divide our experimental results into two different research aspects such as 1) elderly healthcare applications and 2) automatic activity recognition datasets which evaluate the performance of proposed and the state of the art methods.

A. Elderly Healthcare Applications

In this subsection, we evaluate our method by considering three benchmark datasets [15] performed by elderly people at multiple environments (i.e., hospital, home and office). Table I compares the recognition accuracy of proposed and state of the art methods based on same settings as [15].

TABLE I

Recognition Accuracy Comparison between Proposed and State of the Art Methods using Three Healthcare Dataset

Healthcare Applications	Invariant features [30]	Motion features [15]	Proposed method
Smart hospital activities	86.09	90.33	94.82
Smart home activities	88.68	92.33	95.15
Smart office activities	89.43	93.58	95.97

Similarly, we compare our detection activities along with active frames (i.e., frames that contain given activity types) having selected human silhouettes and obtained precision, recall and F-measure. Table II compares the performance of proposed and state of the art methods. It is quite obvious that our method is effective to encode relevant activity information.

TABLE II

Precision, Recall and F-measure Comparisons on the Smart Home Dataset for Activity Localization

Methods	Precision	Recall	F-measure
Invariant features [30]	78.3	80.6	78.9
Motion features [15]	80.9	82.4	81.3
Proposed method	84.7	87.3	85.8

B. Experiments on Automatic Activity Recognition Datasets

In this subsection, we conduct experiments on three challenging depth-based activity and action datasets such as online self-annotated dataset [31], MSRDailyActivity3D and MSRAction3D for recognition purpose using multi-features and embedded HMM. However, the content of each dataset, experimental setting and results are described in the following sub-subsections.

1) Online Self-Annotated Dataset

In this experiment, the depth-video activity dataset [32] is collected for the fifteen different activities based on daily life healthcare monitoring scenarios often encountered by elderly people. These activities include: *sitting down, taking-medicine, falling-down, both hands waving, eating, clapping, phone conversation, walking, exercising, stand up, cleaning, taking an object, reading an article, pointing an object and hand waving,* respectively. All activities are captured in labs and halls. The total dataset consists of 705 video sequences performed by sixteen different subjects.

Its training datasets include 675 segmented video sequences and its testing phase contains 30 unsegmented video sequences having time duration of two to four minutes. From Table III, we illustrate that the proposed method achieved recognition rate of 71.6% for all fifteen activities.

TABLE III

Comparison of Recognition Accuracy between Proposed and State of the Art Methods using Online Self-annotated Dataset

Methods	Accuracy (%)
Dynamic temporal warping [33]	38.7
Multi-part bag-of-poses [34]	47.6
HOJ3D [35]	49.6
Multimodal approach [36]	51.6
Depth silhouettes context features [32]	57.6
Multi-Features method	71.6

While, Table IV summarizes the comparison of our method with the state of the art methods by considering a decision of recognition results based on 100 frame sliding window approach.

TABLE IV

Mean Recognition Accuracy of Proposed Multi-features Method using Online Self-annotated Dataset

Activities	Accuracy (%)	Activities	Accuracy (%)
Sitting down	38.7	Hand waving	67.7
Falling-down	47.6	Taking-medicine	78.1
Eating	51.6	Both hands waving	58.2
Phone conversation	89.9	Clapping	78.5
Exercising	54.8	Walking	81.3
Cleaning	73.4	Stand up	63.7
Reading an article	58.2	Take an Object	69.6
		Pointing an Object	84.4

In addition, Table V shows the performance evaluation of online self-annotated activity recognition dataset from all three indicators such as precision, recall and F-measure. It is clearly justified that the proposed features are significantly better than conventional ones.

TABLE V

Performance Evaluation using Precision, Recall and F-measure Parameters for Online Self-annotated Dataset

Methods	Precision	Recall	F-measure
[30]	53.7	56.2	55.8
[15]	58.3	61.8	60.5
[36]	61.6	62.1	62.8
[32]	63.9	65.3	64.2
Proposed method	68.4	67.9	67.3

2) MSRDailyActivity3D Dataset

The MSRDailyActivity3D dataset [25] was collected with human daily activities captured by the Kinect sensor. This dataset consists of sixteen activities including: drink, eat, read book, call cellphone, write, use laptop, vacuum cleaner, cheer up, sit still, toss paper, play game, lie down, walk, play guitar, stand up and sit down. All subjects perform activities into both standing and sitting on sofa poses. The number of activity video sequences is 320. While, dataset is quite challenging due to human object interactions. Some sample images of MSRDailyActivity3D dataset are shown in Fig. 11.

Fig. 11. Sample depth images used in MSRDailyActivity3D dataset.

However, this dataset includes one sample per subject, therefore, we applied leave-one-subject-out (LOSO) cross validation process in our experiment. Table VI presents the recognition accuracy of our proposed multi-features method.

TABLE VI
Recognition Accuracy Comparison between Proposed and State of the Art Methods using MSRDailyActivity3D Dataset

Methods	Accuracy (%)
Eigenjoints [24]	58.1
Joint position features [25]	68.0
Graph based genetic programming [37]	72.1
Moving Pose[38]	73.8
Integrating Joints features [39]	76.0
Motion features [40]	79.1
Actionlet ensemble[25]	85.7
Super normal vector [41]	86.2
Depth Cuboid Similarity features[42]	88.2
Multi-Features method	92.2

Also, we compare the recognition performance using MSRDailyActivity3D dataset where the proposed method achieved a superior mean recognition rate of 92.2% over the state of the methods [24], [25], [37]-[42] as shown in Table VII.

TABLE VII
Recognition Performance Results of Proposed Multi-features Method using MSRDailyActivity3D Dataset

Activities	Accuracy (%)	Activities	Accuracy (%)
Drink	89.6	Eat	96.2
Read Book	93.4	Call cell phone	97.7
Write	87.5	Use laptop	89.6
Vacuum Cleaner	98.8	Cheer up	96.4
Sit still	87.3	Toss paper	88.1
Play game	89.3	Lie down	98.3
Walk	95.7	Play guitar	87.3
Stand up	90.9	Sit down	89.4

3) MSRAction3D Dataset

The MSRAction3D dataset [20] was captured with a depth sensor (i.e., Kinect device) by the Microsoft Researcher team. It includes 20 different action types as: *high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two hand wave, side boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing and pick up & throw.* The dataset consists of 567 depth map sequences performed by 10 subjects. Also, the background of this dataset is clean and the human silhouettes are available in each frame. This dataset is quite challenging due to similar postures of different actions especially hands and legs movements. Several samples of MSRAction3D dataset are shown in Fig. 12.

Fig. 12. Sample depth images of MSRAction3D dataset.

Also, we follow the same experimental setting as [25] and obtained the recognition accuracy of 93.1% as shown in Table VIII.

TABLE VIII
Recognition Accuracy Comparison using MSRAction3D Dataset

Methods	Accuracy (%)
Dynamic temporal warping [33]	54.0
Bag of 3D points [20]	74.7
HOJ3D [35]	79.0
Motion and Shape features [43]	82.1
Eigenjoints [24]	82.3
Semi Supervised learning [44]	83.5
Grassmannian manifold [45]	86.2
HON4D [21]	88.3
Pose Set [46]	90.0
HOD Descriptor [47]	91.2
Euclidean group algorithm [48]	92.4
Multi-Features method	93.1

In addition, we compare our method with the state of the art methods [31], [20], [35], [24], [21], [43]-[48] on the cross subject test setting and obtained a significantly improved recognition performance over existing works as shown in Table IX.

TABLE IX
Mean Recognition Rate of Proposed Multi-features Method using MSRAction3D Dataset

Activities	Accuracy (%)	Activities	Accuracy (%)
High arm wave	89.5	Horizontal arm wave	90.9
Hammer	98.6	Hand catch	89.8
Forward punch	96.4	High throw	93.6
Draw x	94.1	Draw tick	95.5
Draw circle	98.8	Hand clap	87.7
Two hand wave	97.2	Side boxing	98.8
Bend	98.7	Forward kick	83.9
Side kick	94.1	Jogging	88.7
Tennis swing	88.5	Tennis serve	86.8
Golf swing	93.6	Pick up and throw	97.4

To evaluate the recognition performance based on various codebook sizes, Fig. 13 shows the recognition accuracies of all three depth datasets having different codebook sizes. We determine the codebook size as 128 experimentally using LBG clustering algorithm because the greater codebook size makes minor changes in HAR accuracy.

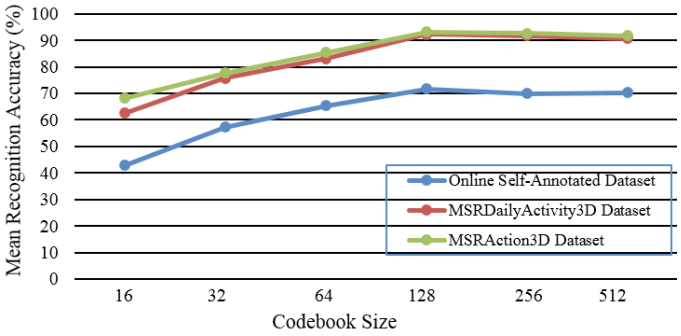

Fig. 13. Recognition accuracies versus different codebook sizes of all three depth datasets.

However, three-state HMM model is selected for the training/testing of all three depth activity/actions datasets after experimenting with different number of states HMMs models as shown in Fig. 14.

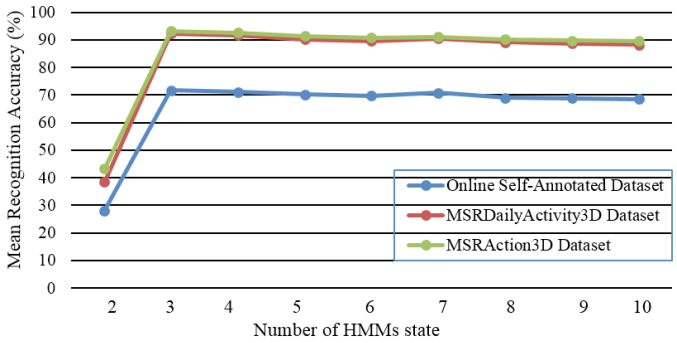

Fig. 14. Recognition accuracies versus different number of states for HMMs using all three depth datasets.

VI. Conclusion

In this paper, a novel approach has been proposed for robust HAR system utilizing multi-features along with embedded HMMs from depth video sensor. The HAR framework contains novel characteristics as (1) a novel real-time body parts tracking system is introduced to extract human silhouettes from noisy background, (2) a robust spatio-temporal multi-features obtained from the full-body human depth silhouettes and joints body parts information, (3) development of new online depth dataset which become a benchmark for video-based HAR systems, and (4) a new concept of embedded HMMs. Our experimental results on three challenging depth datasets have shown the significant recognition performance of our features over the state of the art features extraction techniques. The proposed system is directly applicable to any e-health monitoring systems, such as monitoring healthcare problems for elderly and sick people, or examines the indoor activities of people at home or hospitals.

In the future work, we will exploit the effectiveness of our features by merging the RGB features along with multi-view invariant characteristics over more complex activities datasets including human-to-human interactions and human-object interactions. Also, some discriminative/generative models are used to robust training/recognition phase to strengthen our HAR algorithm.

References

[1] P. Petersen, D. Kandelman, S. Arpin, and H. Ogawa, "Global oral health of older people-Call for public health action," *Community dental health*, vol. 27, no. 4, pp. 257–268, Dec. 2010.

[2] V. Osmani, S. Balasubramaniam, and D. Botvich, "Human activity recognition in pervasive health-care: Supporting efficient remote collaboration," *Journal of Network and Computer Applications*, vol. 31, no.4, pp. 628–655, Nov. 2008.

[3] J. Dunn, M. Rudberg, S. Furner, and C. Cassel, "Mortality, disability, and falls in older persons: the role of underlying disease and disability," *American Journal of Public Health*, vol. 82, no. 3, pp. 395–400, Mar. 1992.

[4] U. Reinhardt, "Does the aging of the population really drive the demand for health care?," *Health Affairs*, vol. 22, no. 6, pp. 27–39, Nov. 2003.

[5] E. Mynatt and W. Rogers, "Developing technology to support the functional independence of older adults," *Ageing International,* vol. 27, no. 1, pp. 24–41, Dec. 2001.

[6] A. Jalal, Y. Kim, Y. Kim, S. Kamal, and D. Kim, "Robust human activity recognition from depth video using spatiotemporal multi-fused features," *Pattern recognition*, vol. 61, pp. 295–308, Jan. 2017.

[7] W. Xu, M. Zhang, A. Sawchuk, and M. Sarrafzadeh, Co-recognition of human activity and sensor location via compressed sensing in wearable body sensor networks, in *Proc. 2012 Ninth international Conf. Wearable and Implantable Body Sensor Networks*, London, 2012, pp. 124–129.

[8] M. Kreil, B. Sick, and P. Lukowicz, "Dealing with human variability in motion based, wearable activity recognition," in *Proc. IEEE international Conf. Pervasive Computing and Communications Workshops*, Budapest, 2014, pp. 36–40.

[9] S. Kamal and A. Jalal, "A hybrid feature extraction approach for human detection, tracking and activity recognition using depth sensors," *Arabian Journal for Science and Engineering*, vol.41, no. 3, pp. 1043–1051, Mar. 2016.

[10] S. Kamal, A. Jalal and D. Kim, "Depth Images-based Human Detection, Tracking and Activity Recognition Using Spatiotemporal Features and Modified HMM," *Journal of electrical engineering and technology*, vol.11, no. 6, pp. 1857–1862, Nov. 2016.

[11] D. Chen, A. Bharucha, and H. Wactlar, "Intelligent video monitoring to improve safety of older persons," in *Proc. IEEE Conf. Engineering in Medicine and Biology Society*, Lyon, 2007, pp. 3814–3817.

[12] A. Jalal, N. Sharif, J. Kim, and T. Kim, "Human activity recognition via recognized body parts of human depth silhouettes for residents monitoring services at smart homes," *Indoor and Built Environment*, vol.22, pp. 271–279, Aug. 2013.

[13] A. Jalal, S. Kamal, and D. Kim, "Depth Map-based Human Activity Tracking and Recognition Using Body Joints Features and Self-Organized Map," in *Proc. IEEE Conf. on computing, communication and networking technologies*, China, 2014, pp. 1–6.

[14] D. Sanchez, M. Tentori, and J. Favela, "Activity recognition for the smart hospital," *IEEE Intelligent systems*, vol. 23, no.2, pp. 50–57, Mar. 2008.

[15] A. Farooq, A. Jalal, and S. Kamal, "Dense RGB-D map-based human tracking and activity recognition using skin joints features and self-organizing map," *KSII Transactions on internet and information systems*, vol. 9(5), pp. 1856-1869, 2015.

[16] A. Jalal, S. Kamal, and D. Kim, "A depth video sensor-based life-logging human activity recognition system for elderly care in smart indoor environments," *Sensors*, vol. 14(7), pp. 11735-11759, 2014.

[17] M. Raptis and L. Sigal, "Poselet key-framing: A model for human activity recognition," in *Proc. IEEE Conf. Computer vision and pattern recognition*, Oregon, 2013, pp. 2650–2657.

[18] N. Cuntoon and R. Chellappa, "Epitomic representation of human activities," in *Proc. IEEE Conf. Computer vision and pattern recognition*, Minneapolis, 2007, pp. 1–8.

[19] A. Bobick and J. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 23, no. 3, pp. 257–267, Mar. 2001.

[20] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *Proc. IEEE Conf. Computer vision and pattern recognition workshops*, San Francisco, 2010, pp. 9–14.

[21] O. Oreifej and Z. Liu, "Hon4d: Histogram of oriented 4d normal for activity recognition from depth sequences," in Proc. *IEEE Conf. Computer vision and pattern recognition*, Oregon, 2013, pp. 716–723.

[22] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu, "Robust 3d action recognition with random occupancy patterns," in *Proc. European Conf. Computer Vision*, Florence, 2012, pp. 872–885.

[23] J. Sung, C. Ponce, B. Selman and A. Saxena, "Unstructured human

activity detection from rgbd images," in *Proc. IEEE conf. Robotics and Automation*, Saint Paul, 2012, pp. 842–849.

[24] X. Yang and Y. Tian, "Eigenjoints-based action recognition using naive-bayes-nearest-neighbor," in *Proc. IEEE Conf. Computer vision and pattern recognition workshops*, Providence, 2012, pp. 14–19.

[25] J. Wang, Z. Liu Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. IEEE Conf. Computer vision and pattern recognition*, Providence, 2012, pp. 1290–1297.

[26] A. Jalal and Y. Kim, "Dense depth maps-based human pose tracking and recognition in dynamic scenes using ridge data," in *Proc. IEEE Conf. Advanced video and signal based surveillance*, Seoul, 2014, pp. 119–124.

[27] A. Jalal, S. Kamal, and D. Kim, "Human depth sensors-based activity recognition using spatiotemporal features and hidden markov model for smart environments," *Journal of computer networks and communications*, vol. 2016, pp. 1–11, Sep. 2016.

[28] V. Deepu, S. Madhvanath, and A. Ramakrishnan, "Principal component analysis for online handwritten character recognition," in *Proc. IEEE Conf. Pattern recognition*, Cambridge, 2004, pp. 327–330.

[29] A. Jalal, Y. Kim, S. Kamal, A. Farooq and D. Kim, "Human daily activity recognition with joints plus body features representation using Kinect sensor," *in Proc. IEEE conference on Informatics, electronics and vision*, Japan, 2015, pp. 1–6.

[30] Y. Wang, K. Huang and T. Tan, "Abnormal activity recognition in office based on R transform," *in Proc. IEEE conference on image processing*, San Antonio, 2007, pp. 341–344.

[31] A. Jalal, "IM-DailyDepthActivity dataset," imlab.postech.ac.kr/ databases. htm, 2015, [Online; accessed 19 August- 2016].

[32] A. Jalal, S. Kamal and D. Kim, "Individual detection-tracking -recognition using depth activity images," *in Proc. IEEE conference on Ubiquitous robots and ambient intelligence*, Korea, 2015, pp. 450–455.

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

[34] L. Seidenari, V. Varano, S. Berretti, A. Bimbo, and P. Pala, "Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses," in *Proc. IEEE Conf. computer vision and pattern recognition workshops*, Portland, 2013, pp. 479–485.

[35] L. Xia, C. Chen, and J. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *Proc. IEEE Conf. Computer vision and pattern recognition workshops*, Providence, 2012, pp. 20–27.

[36] A. Keceli and A. Can, "A multimodal approach for recognizing human actions using depth information," in *Proc. IEEE Conf. Pattern recognition workshops*, Stockholm, 2014, pp. 421–426.

[37] L. Liu and L. Shao, "Learning discriminative representations from RGB-D video data," in *Proc. Inter. Conf. on artificial intelligence*, Beijing, 2013, pp. 1493–1500.

[38] M. Zanfir, M. Leordeanu, and C. Sminchisescu, "The moving pose: an efficient 3d kinematics descriptor for low-latency action recognition and detection," in *Proc. IEEE Conf. Computer Vision*, Sydney, 2013, pp. 2752–2759.

[39] Q. Li, Y. Zhou and A. Ming, "Integrating joint and surface for human action recognition in indoor environments, in *Proc. IEEE Conf. security, pattern analysis and cybernetics*, China, 2014, pp. 100–104.

[40] A. Jalal and S. Kamal, "Real-Time Life Logging via a Depth Silhouette-based Human Activity Recognition System for Smart Home Services," *in Proc. of the IEEE International Conference on Advanced Video and Signal-based Surveillance*, Korea, 2014, pp. 74-80.

[41] X. Yang and Y. Tian, "Super normal vector for activity recognition using depth sequences, in *Proc. IEEE Conf. Computer vision and pattern recognition*, Columbus, 2014, pp. 804–811.

[42] L. Xia and J. Aggarwal, "Spatio-temporal Depth cuboid similarity feature for activity recognition using depth camera, in *Proc. IEEE Conf. Computer vision and pattern recognition*, Portland, 2013, pp. 2834–2841.

[43] A. Jalal, S. Kamal and D. Kim, "Shape and motion features approach for activity tracking and recognition from Kinect video camera, in *Proc. IEEE Conf. advanced information networking and applications workshops*, Korea, 2015, pp. 445–450.

[44] M. Mabrouk, N. Ghanem and M. Ismail, "Semi supervised learning for human activity recognition using depth cameras, in *Proc. IEEE Conf. on machine learning and applications*, US, 2015, pp. 681–686.

[45] R. Slama, H. Wannous and M. Daoudi, "Grassmannian representation of motion depth for 3D human gesture and action recognition, in *Proc. Inter. Conf. on Pattern recognition*, Sweden, 2014, pp. 3499–3504.

[46] C. Wang, Y. Wang, and A. Yuille, "An approach to pose based action recognition," in *Proc. IEEE Conf. Computer vision and pattern recognition*, Portland, 2013, pp. 915–922.

[47] M. A. Gowayyed, M. Torki, M. E. Hussein and M. El-Saban, "Histogram of oriented displacements (HOD): describing trajectories of human joints for action recognition," in *Proc. Inter. Conf. on artificial intelligence*, China, 2013, pp. 1351–1357.

[48] R. Vemulapalli, F. Arrate and R. Chellappa, "Human action recognition by representing 3D skeletons as points in a lie group," in *Proc. IEEE Conf. computer vision and pattern recognition*, Columbus, 2014, pp. 588–595.

Ahmad Jalal received his B.S. degree in Computer Science from Iqra University, Peshawar, Pakistan and M.S degree in Computer Science from Kyungpook National University, Republic of Korea. He received his Ph.D. degree in the Department of Biomedical Engineering at Kyung Hee University, Republic of Korea. Now, he is working as Post-doctoral Research fellowship at POSTECH (Pohang University of Science and Technology). His research interest includes human computer interaction, image processing and computer vision.

Shaharyar Kamal He received his B.S. degree in Software Engineering from City University of Science and Information Technology, Peshawar, Pakistan and M.S. degree in Computer Engineering from Mid Sweden University, Sweden. He is currently enrolled in Ph.D. in the Department of Radio and Electronics Engineering at Kyung Hee University, Republic of Korea. His research interest includes wireless communication, image and signal processing.

Daijin Kim He received the B.S. degree in electronic engineering from Yonsei University, Seoul, Korea, in 1981, and the M.S. degree in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Taejon, 1984. In 1991, he received the Ph.D. degree in electrical and computer engineering from Syracuse University, Syracuse, NY. During 1992–1999, he was an Associate Professor in the Department of Computer Engineering at Donga University, Pusan, Korea. He is currently a Professor in the Department of Computer Science and Engineering at POSTECH, Pohang, Korea, a vice president of office of academic information affairs and Director of BK21 þ POSTECH CSE Institute. His research interests include computer vision, human computer interaction, and intelligent systems.