Recognizing Involuntary Actions from 3D Skeleton Data Using Body States

Mozhgan Mokari, Hoda Mohammadzade*, Benyamin Ghojogh

***Abstract*—Human action recognition has been one of the most active fields of research in computer vision over the last years. Two dimensional action recognition methods are facing serious challenges such as occlusion and missing the third dimension of data. Development of depth sensors has made it feasible to track positions of human body joints over time. This paper proposes a novel method for action recognition which uses temporal 3D skeletal Kinect data. This method introduces the definition of body states and then every action is modeled as a sequence of these states. The learning stage uses Fisher Linear Discriminant Analysis (LDA) to construct discriminant feature space for discriminating the body states. Moreover, this paper suggests the use of the Mahalonobis distance as an appropriate distance metric for the classification of the states of involuntary actions. Hidden Markov Model (HMM) is then used to model the temporal transition between the body states in each action. According to the results, this method significantly outperforms other popular methods, with recognition (recall) rate of 88.64% for eight different actions and up to 96.18% for classifying the class of all fall actions versus normal actions.**

***Index Terms*—Human action recognition, involuntary action recognition, Fisher, linear discriminant analysis (LDA), kinect, 3D skeleton data, hidden markov model (HMM).**

I. Introduction

SINCE last two decades, human action recognition has drawn lots of attention from researches in computer vision and machine learning fields. In early attempts for action recognition, RGB video was used as input of recognition system. Various valuable methods and algorithms were proposed for recognizing actions and activities using RGB data. However, several problems exist in action recognition using RGB frames such as occlusion and different orientations of camera. Existence of other objects in addition to human bodies and the lack of information of the third dimension can be mentioned as other challenges in this category of methods [45], [12], [41], [24], [24], [40], [30]. In order to address these problems, methods for recognizing action from multiple views have been also introduced; however, they are typically very expensive in calculations and are not suitable for real time recognition [15].

Due to the mentioned problems and by introducing 3D Kinect sensors in market, researchers started to work on 3D data for the purpose of action recognition. The Kinect sensor provides both depth and skeleton data in addition to capturing RGB frames. Different methods have been proposed so far for action recognition using either depth or skeleton data.

Action recognition has a variety of different applications. From one point of view, all actions can be categorized in one of the two categories of normal (voluntary) and involuntary actions (see Fig. 1). Daily actions, actions for gaming, and interactions between human and computer can be considered as normal actions. On the other hand, involuntary actions can happen in different places, such as homes, hospitals and public places. One of the most frequent involuntary actions is falling which can happen by patients in hospitals. Old people are also subject to dangerous falls, which if detected by surveillance systems for elderly cares can reduce serious injuries and fatalities. Another example where proper detection of involuntary actions can prevent problems and chaos is in public places. In these places, involuntary actions such as falling or being thrown can happen as a result of accident or physical fight. In comparison to normal actions, involuntary actions usually have larger performance variance among various trails and different subjects. This characteristic of involuntary actions is the main challenge in recognizing them. Although the proposed method in this work can be applied for both normal and involuntary actions, its focus is on involuntary actions and tries to handle the mentioned challenge. Figure 2 depicts a human action recognition system used for fall detection. As it is seen, it is not possible to train the system using all various types of fall actions over all different subjects. Therefore, the challenge is to recognize any fall action using limited number of training samples.

This paper proposes a new method for human action recognition, especially for involuntary actions. The main contributions are as follows:

- In contrast to most of action recognition methods in literature, this work is not feature-based but is holistic. In other words, features (such as histogram of joints as used in [44]) are not extracted from skeleton but the raw features of skeletons are fed to the so called body state classifier. Consequently, the classifier is responsible to extract discriminant features. As it is well-known in face recognition [48], holistic methods have more potential for accurate recognition because of using all the information and devolving feature extraction to classifier. Our experiments verify our better performance in comparison to feature-based methods, such as [44], in both action-vs-action and normal-vs-fall actions scenarios.
- This work properly handles involuntary actions, which are variously distributed in the space of joints, by taking into account the distribution for each body state.

Mozhgan Mokari's e-mail: mozhgan.mokari@ee.sharif.ir
Hoda Mohammadzade's e-mail: hoda@sharif.edu (corresponding author)
Benyamin Ghojogh's e-mail: ghojogh_benyamin@ee.sharif.edu
All authors are with Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran

- Different speeds in performing involuntary actions are handled by using Hidden Markov Models (HMM).
- This method can be used for recognizing normal actions as well as involuntary ones.
- Other than outperforming in recognition of each of the various normal and involuntary actions in the dataset, the proposed method achieves a great recognition rate for classifying the class of all involuntary actions versus normal actions. This scenario is particularly important where only the involuntary action detection is important, such as elderly or patient surveillance.

This paper is organized as follows. Section II reviews related work. Section III proposes the main algorithm of proposed method which includes modeling human body, and action recognition using Fisher Linear Discriminant Analysis (LDA) and Hidden Markov Model (HMM). Section IV introduces the utilized dataset and experimental results. Finally, Section V concludes the article and addresses the possible future work.

II. Related Work

According to the importance of action recognition and its large amount of applications, lots of different methods have been proposed in this field. In [32], Peng et al. described different kinds of bag of visual words model (BoVW) methods and investigated the effect of each of them on action recognition. These factors were feature extraction, feature preprocessing, codebook generation, feature encoding, pooling, normalization, and fusing these descriptors.

Liu et al. [26] employed Genetic Programming (GP) on spatio-temporal motion features for action recognition. Features were extracted from both color and optical flow sequences. Wang et al. [42] used homography for cancellation of camera motions from trajectories and optical flows. SURF descriptors and dense optical flows were employed with RANSAC to estimate this homography. Then, motion-based histogram of optical flows (HOF) and motion-based histogram (MBH) descriptors were used for action recognition.

Facing some challenges such as coverage of some part of body by others and introducing 3D methods, encouraged researchers to use depth map. Li et al. [21] recognized human's action by sampling the 3D points of depth image and creating an action graph. In this method, they modeled the position of human's body by projecting the contour of body shape onto the different planes and sampling them. Then, the state of human body was modeled with these bags of 3D points. The states were considered as nodes of a graph, modeling the action. Although this method is not robust to the changing of viewing angle and human's body scale, it has recognized 90 percent of actions and the error was halved compared to 2D methods. Rahmani et al. [34] used Histogram of Oriented Principal Components (HOPC) descriptor on point clouds for cross-view action recognition.

Zhao et al. [49] classified human's actions by utilizing information of RGB and depth image. They obtained spatio-temporal interest points from RGB image and used combined descriptor of RGB and depth images.

Liu et al. [27] encoded sptial-temporal information of skeleton joints in depth sequences into color images. In this regard, 5D space of (x, y, z, f, n) was expressed as a 2D coordinate space and a 3D color space, where f and n denote time and joint labels, respectively. Convolutional neural network was used to extract more discriminative deep features. These features were used for action recognition.

Rahmani and Mian [35] transferred human poses to a view-invariant high-level space and recognized action in depth image by using deep convolutional neural network. Their method obtained appropriate results in multi-view datasets. In [47], Zhang et al. used 3D Histograms of Texture (3DHoTs) from depth maps. The 3DHoTs were formed by characterizing the salient information of action. In their method, action was represented by texture features. classification of actions was done by multi-class boosting classifier (MBC).

Chen et al. [7] projected depth videos on three orthogonal Cartesian planes. Absolute difference between two consecutive projections was accumulated creating Depth Motion Maps (DMMs). Then action recognition was performed by distance-weighted Tikhonov matrix with an I2-regularized classifier. Chen et al. [8] proposed a Local Binary Patterns (LBP) descriptor which is invariant to shape and speed for action recognition in depth videos. They partitioned Depth Motion Maps (DMMs) and extracted LBP for action recognition. Liang et al. [22] applied DMMs-based Gradient Local Auto-Correlations (GLAC) features of depth videos to capture the shape information of sub-actions. They proposed Locality-constrained Affine Subspace Coding (LASC) to encode the extracted features. This method had competitive results with less complexity.

By developing Kinect sensors and related software for tracking humans in images and detecting positions of body joints in 3D space, several methods were proposed to recognize action using this information. One of these methods introduced Cov3DJ descriptor [16] which separated different action classes by finding covariance matrix of positions of the joints during the action, and used Support Vector Machine (SVM) for classification.

Reddy et al. [36] recognized action by considering mean, minimum and maximum of position of joints as features and compared them to features obtained by using Principle Component Analysis (PCA) on position of joints. Likewise, Martínez-Zarzuela et al. [29] tried to recognize actions by taking a sequence of positions of joints as a signal and extracting the five first Fast Fourier Transform (FFT) components as a feature vector which was fed into a neural network. However, this method did not perform very well for complex actions involving different parts of body.

As different actions involve different joints, Anjum et al. [5] selected important and effective joints in the training level, according to the type of action. In their method, each action was determined by three joints. Results showed that this method performs better with less information but joints should be selected in training for each action. Therefore, extending this algorithm for new actions is time-consuming and expensive.

Liu et al. [25] used tree-structure based traversal method on 3D skeleton data and extended RNN-based learning method to spatio-temporal domain. In this way, they could analyze

the hidden sources of information in actions. Ke et al. [19] transformed skeleton sequences into clips consisting spatial temporal features. They used deep convolutional neural networks to learn long-term temporal information. Multi-Task Learning Network (MTLN) was used to incorporate spatial structural information for action recognition.

In [37], Shahroudy et al. described actions by partitioning kinetics of body parts. They used sparse set of body part to model action as a combination of multimodal features. Dynamics and appearance of parts were represented by heterogeneous set of depth and skeleton-based features. Huynh et al. [17] proposed a new method more robust to human scale and changes of position. They categorized joints into three classes of stable, active and highly active joints, and utilized angles of 10 important joints and vectors connecting moving joints to stable joints. Their method performed better than a similar method which uses only raw position of joints.

Luvizon et al. [28] selected subgroups of joints by Vector of Locally Aggregated Descriptors (VLAD) algorithm. Classification accuracy was improved by the non-parametric K-NN classifier with Large Margin Nearest Neighbor (LMNN). Amor et al. [4] used skeleton shapes as trajectories on Kendall's shape manifold to represent the special dynamical skeletons.

Xia et al. [44] used middle and side hip joints to extract a histogram of position of other joints to be used as feature vector. They reduced the dimension of the feature vector using Linear Discriminant Analysis (LDA) and used K-means method to cluster the feature vectors. Each cluster constituted a visual word. Each action was determined as a time sequence of these visual words and modeled by Hidden Markov Model (HMM). Results showed that this method partially overcame challenges such as different lengths of actions and the same action done in different ways and view angles.

Papadopoulos et al. [31] obtained orientation of body using the positions of shoulders and hip joints and thereby, extracted orthogonal basis vectors for each frame. A new space was then constructed for every person according to its orientation of body. According to these vectors and the new space, the spherical angles of joints were used instead of position of joints. The use of angles instead of position of joints, made the method more robust against human's body scale and changes in the shape of body. This method also used energy function to overcome the challenge of same actions done by opposite hands or feet.

Although there are lots of proposed methods for action recognition, but many problems and challenges still remain unsolved. This paper tries to tackle some of them such as different distributions of actions in statistical feature space, especially for involuntary actions.

III. Methodology

In order to recognize actions, at the first step, the actions should be modeled in an appropriate way. Modeling actions depends on various facts such as application, types of actions and method of classification. One of the most important applications of action recognition is online recognition where the recognition should be performed in real time. This article goals this type of recognition. In this category, the action should be modeled so that the model can be updated during completion of action and finally recognize the type of performed action. Therefore, in this article, each action is supposed to be a sequence composed of several states of body.

In the next step, position of joints in the 3D space are utilized in order to model the state of body. The position of joints are prepared by the output of Kinect sensor. The skeleton consists of several joints, which are 25 joints for the dataset used for the experiments in this paper. A number of these joints are, however, very close to each other without any important difference in movements; therefore, their information are almost redundant. With respect to the actions addressed in this paper, merely 12 important joints, which are right and left ankles, right and left knees, right and left wrists, right and left shoulders, head, middle spine, hip and spine shoulder are selected out of the skeleton. Position of spine base (hip) and right and left shoulders are used for alignment in order to correctly describe the state of body in different persons and runs. The selected joints and also joints required for alignment are shown in Fig. 3. State modeling including skeleton alignment and state classification are detailed in the following.

A. Modeling State of Body

In order to model and describe the state of body, a proper descriptor should be created. This descriptor models the action as a time sequence of states and tries to recognize the action. The body states are determined as follows. According to nature of every action, the main body states, of which the action is composed, are conjectured and then are manually selected and sampled out of the training sequences of frames. Notice that this manual selection is done merely in training phase, while in the test phase, each input frame is automatically classified by the classifier of body states.

1) Aligning Skeleton: Different locations and orientations of body in the frames forces the need to aligning the skeleton. As already mentioned, 12 joints positions are used in 3D space, in order to describe the state of body. In order to cancel the location of body skeleton, the position of hip joint is subtracted from the position of all joints. This is performed for every frame in the sequence.

Moreover, different orientations of skeleton or camera makes recognizing similar states difficult and even wrong. Thus, in order to cancel different orientations of body skeletons, the body is rotated around y axis making the projection of the vector connecting the left and right shoulder onto the xz plane parallel to the x axis. By performing this rotation, the skeleton directly faces the camera. This procedure is illustrated in Fig. 4. The axes can be seen in Fig. 3. In literature, the alignment of skeleton is often performed, but the methods or the joints used for that might differ. For example, in [44], left and right hip joints are utilized rather than shoulder joints for alignment.

2) Creating feature vector: To determine the state of body in each frame, proper feature vectors are required. Three

joints out of the 12 joints are used for alignment and the remaining nine joints are used to create the feature vectors. If (x_m, y_m, z_m) denotes the coordinate of m^{th} joint ($m = \{1, \ldots, 9\}$), the raw feature vector are obtained as $[x_1, \ldots, x_9, y_1, \ldots, y_9, z_1, \ldots, z_9]^T$. Fisher Linear Discriminant Analysis (LDA) [10], [14] is utilized for extracting discriminant features from the raw feature vectors. In Fisher LDA method, the dimension of feature vector is reduced to $C-1$, where C is the number of states. In LDA, the within- (S_w) and between-class (S_b) scatter matrices are

$$S_w = \sum_{i=1}^{C} \sum_{x_k \in X_i} (x_k - \mu_i)(x_k - \mu_i)^T, \tag{1}$$

$$S_b = \sum_{i=1}^{C} N_i(\mu_i - \mu)(\mu_i - \mu)^T, \tag{2}$$

in order to minimize the within class covariance and maximize the between class covariance [14], [6], where μ_i and μ denote the mean of i^{th} state and the mean of class means, respectively. The Fisher projection space is created by the eigenvectors of $S_w^{-1}S_b$. By its projection in this space, the feature vector F for an input skeleton state is obtained.

After projection onto Fisher space, the obtained feature vectors are located relative to each other such that those relating to similar and different states, respectively fall close and apart. By this fact, recognition of states becomes available.

There are also other methods for feature reduction which can be used for classification. One of the most popular methods of this category is Principle Component Analysis (PCA) [14], [6]. However, PCA method cannot always classify the data as well as LDA does. As an example, suppose that the distribution of classes are similar to that depicted in Fig. 5. In this example, the Fisher LDA direction is perpendicular to the direction of PCA. As is obvious in this figure, Fisher LDA tries to minimize within-class variance while maximizes between-class variance in order to classify them.

The resulting feature vectors are used for training and testing the state of body. The action will be defined as a time sequence of multiple specific states. The state of body is recognized in the test phase, by finding the minimum distance as described in the following section.

3) Finding the minimum distance: In every frame denoted as f, the state of body should be recognized. For reaching this goal, the distances between feature vector F of this frame and the means of the feature vectors of all states are found. The minimum distance determines the state of the frame f. If $\widetilde{F}_i$ denotes the mean of feature vectors of the i^{th} class, the state is found as,

$$\text{state}(f) = \arg\min_i d(F, \widetilde{F}_i), \tag{3}$$

where d is the distance measurement function which can be one of the two followings:

- *Euclidean Distance:* One of the most popular methods for calculating the distance of two vectors is Euclidean distance, which is used as one of the distance methods in this article. The function of Euclidean distance can be formulated as,

$$d(F, \widetilde{F}_i) = \sqrt{\sum_j (F_j - \widetilde{F}_{ij})^2}, \tag{4}$$

where F_j and $\widetilde{F}_{ij}$ are the j^{th} components of F and $\widetilde{F}_i$, respectively.

- *Mahalanobis Distance:* As the minimum distance from the means of states is used for recognizing the state, using a proper distance has much important influence on the accuracy of recognition. Therefore, the distribution of final feature vectors in the feature space should be considered and the distance measurement should be defined accordingly.

If body states are categorized into C classes, the dimension of the final feature (Fisher) vectors would be $C-1$. As the dimension of the feature vectors might be high, their distribution in each class cannot be directly visualized for direct analysis. However, the distribution of feature vectors can be analyzed in higher dimensions by calculating their covariance matrices. The first two directions of Fisher space are used here for illustrating the distribution of each of the eight body states defined for the TST dataset, which are discussed in more details in Section IV. Figure. 6 illustrates the training samples projected onto the space constructed by the first two Fisher directions. As shown in this figure, distribution of feature vectors for each state is different in the two directions.

The more differently people perform an action containing a state, the wider the distribution for the state would be. The more widely distributed states are usually those during the completion of an involuntary action. For instance, as shown in Fig. 6, after projection on constructed Fisher space, states related to normal actions such as standing and sit states are less distributed than the states occurred in involuntary actions, such as lay front and lay back. In order to handle the challenge of different distributions of projected states, a distance measurement function other than Euclidean one should be used which considers the distributions.

Mahalanobis distance considers variances of distributions in its calculation, which is calculated as,

$$d(F, \widetilde{F}_i) = \sqrt{(F - \widetilde{F}_i)^T S^{-1} (F - \widetilde{F}_i)}, \tag{5}$$

where S denotes the covariance matrix of the feature vectors of the class to which the distance is calculated. As is obvious in equation (5), the covariance matrix S acts as a weighting matrix for each class according to its distribution. That is, the importance of distance in a particular dimension is considered in calculating the distances. In other words, the distance in a direction with smaller variance is less valuable, yielding to S^{-1} in the equation.

Mahalanobis distance is actually an extension to the standard deviation from the mean, in multi-dimensional space. Experiments reported in the following sections,

show outperformance of this distance in comparison with Euclidean distance.

B. Classyfing Actions Using Hidden Markov Model

As previously mentioned, every action can be modeled as a sequence of consequent states. After recognizing states of body using Fisher LDA, Hidden Markov Model (HMM) is utilized in this work to classify actions.

Every action is modeled using a separate HMM. Each HMM has a number of hidden states with specific transition probabilities between them. For instance, a three-state HMM and its transition probabilities are illustrated in Fig. 7 [39]. Every hidden state has specific emission probabilities for emitting body states. The transition and emission probabilities of each HMM are estimated by the well-known Baum-Welch expectation maximization algorithm [33] using the training observations, i.e., sequences of body states. This algorithm starts with initial assumptions for all of the parameters of the model (i.e., transition and emission probabilities) and then updates the parameters using corresponding expectation maximization equations iteratively until convergence.

In order to decrease computational cost of the algorithm, the frame per second rate has been reduced by down sampling. Uniform down sampling with the rate of 20 is used which was shown appropriate according to our experiments [1]. After constructing a HMM for each action, an unknown sequence is recognized by feeding it to each HMM. After feeding the test sequence of frames to all trained HMMs, every HMM outputs a probability of occurrence for that sequence. The maximum probability determines the action of that sequence.

To have more intuition, note that every period of repetitions of a body state can be roughly associated to a HMM state. For example, when having three-states HMMs for classifying actions, the actions sit, grasp, and end up sit are mostly made of the sequences $\{\overline{\text{standing}}, \overline{\text{crouching}}, \overline{\text{sit on chair}}\}$, $\{\overline{\text{standing}}, \overline{\text{bend}}, \overline{\text{standing}}\}$, and $\{\overline{\text{standing}}, \overline{\text{crouching}}, \overline{\text{sit on ground}}\}$, where $\overline{\text{body state}}$ denotes a sub-sequence of repetitions of the body state (more details about how body states are defined are discussed in Section IV). Moreover, in each sub-sequence, the number of repetitions of the corresponding body state can be different across subjects and different trials.

For each action, the sequences that are used for training HMM are adjusted to have the same lengths (number of body states). This equalization is performed by manually repeating the last state done by the person; so that the total number of states of all actions become equal. It is important to note that this equalization does not compensate for the different lengths and speeds of actions performed by different people or over different trials.

The advantage of HMM, in this work, is that it considers solely the dynamic of sequence and is not sensitive to various paces and lengths of actions. For instance, there exist sequences of lengths 75 frames upto 463 frames with different speeds of actions in TST fall dataset [1], [11], and these sequences have been successfully handled and recognized by this method.

The overall structure of the proposed framework is summarized in Fig. 8.

[1]Note that the sampling rate of the Kinect V2 sensor is known to be 30 frames per second (fps) in normal lighting conditions and 15 fps in poor lighting conditions. According to the corresponding RGB images of the dataset, the samples in this dataset should have been captured in normal lighting condition, and hence, the original sampling rate for this dataset must be 30 fps.

IV. Experimental Results

To examine the proposed method, TST Fall Detection dataset [1] is used. The details of this dataset are explained in next section followed by the explanation on how the actions are modeled in this dataset. At the end, the results of the experiments are presented.

A. Dataset

TST Fall Detection dataset [1], [11] is used for verifying the effectiveness of this method. There are two main categories of actions in this dataset, i.e., daily living activities and fall actions. 11 different persons perform every action for three times. The daily living activities are sit, lay, grasp and walk and the fall actions are falling front, back, side and end up sit.

This dataset has prepared information of 3D position of joints and depth data obtained by the Kinect sensor V2, which is more accurate than previous Kinect sensors. Only the skeletal data of this dataset is used in this work for experiments.

As previously mentioned, one of the important goals in human action recognition is surveillance application especially for controlling elderly or patient people. The main goal of detecting involuntary actions and improvements of Kinect V2 encouraged this work to use the mentioned dataset. Unlike other datasets, involuntary actions, such as falling down exist sufficiently in this dataset, which makes this database challenging.

As fall actions are performed involuntarily, different states and conditions from normal actions appear in different people. Therefore, existing action recognition methods may not necessarily perform as well for fall actions. Moreover, a number of methods have been proposed to recognize fall actions, which concentrate on using features such as speed and acceleration recorded by accelerometer sensors. These features are not able to effectively discriminate the normal actions from each other nor involuntary actions from each other, and therefore do not help recognizing the actions in general. Therefore, the main challenge here is to develop a method which can detect and analyze both of the normal and involuntary actions and also recognize them from each other.

Several samples of depth images of actions in TST dataset are shown in Fig. 9.

B. Recognition of states

In the dataset, only the actions are labeled, and therefore labeling states should be performed manually. According to the actions, eight different states are chosen and labeled to be used to train and test the state classification module. The

chosen states should include the main states of actions in the dataset and should not contain unnecessary states which are close to other states. The chosen states are standing, crouching, lay back, lay front, lay side, bend, sit on chair and sit on ground. An example of each state is shown in Fig. 10.

The "leave one subject out" cross validation is used for the experiments. In each iteration, the entire samples of a person is considered as test samples and the samples of other subjects are used for training system. This type of cross validation is fairly difficult because the system does not see any sample from the test subject in training phase. The state recognition experiment is repeated using both of the distance methods, and the results are listed in Table I. Note that all the rates reported in this paper are recall rates ($\frac{\text{true positive}}{\text{true positive+false negative}}$), unless where the type of rate is mentioned.

Table I shows that the Mahalanobis distance outperforms the Euclidean distance in general. As was expected, the recognition rates of crouching, lay front and bend have been improved significantly using Mahalanobis distance. The reason is that the variances of training data for these states are huge and this fact is not taken into account when Euclidean distance is used.

It is worth noting that using the Mahalanobis distance, the recognition rate of bend state has been improved at the cost of reducing the recognition rate of standing state. Closer look at Fig. 6 reveals that there exists an overlapping region of distributions between the two states. Euclidean distance which does not consider the distribution of classes, mostly recognizes the overlapping region as the standing state. On the other hand, the Mahalanobis distance mostly recognizes this region as the bend state because the variance of standing state is much less than bend. This fact can also be seen from the confusion matrices of states for both distances which are depicted in Fig. 11.

C. Action Recognition & Comparison

In the last step, an action is represented as a sequence of states. Each state in this sequence is recognized by projecting into the LDA space and utilizing a distance measure. Then, the probability of each HMM (note that there is an HMM for each specific action) generating the input sequence of states is calculated and maximum probability determines the recognized action. The number of hidden states in HMM's (note that hidden states are different from body states) affects the recognition performance. Therefore, different number of hidden states were tested for HMM's in this work and were compared to each other. Results of three different numbers of hidden states for HMM's are reported in Table II. The experiments of this table are performed with Mahalanobis distance. As was expected according to the nature of states and actions in the TST Fall dataset [1], [11], HMM's with three hidden states perform better and hence, the number of hidden states for HMM's is considered to be three in this work. It is worth nothing that, combination of optimum number of hidden states for each action was also considered, but the experiments showed that use of a constant number of hidden states for all HMM's results in better performance.

In this article, the proposed method is compared with the method of Xia et al. [44] which has received considerable attention in literature [2], [13], [9], [46] and has been used for comparison in very recent methods [43], [3], [38], [18], [23]. Note that all the above methods has experimented with datasets created using an older version of Kinect sensor and not containing involuntary actions.

For implementing method [44] and fairly comparing it with the proposed method using the TST dataset, several necessary adjustments were performed in its settings. First, for LDA, the states are labeled in the same way as in the proposed method. Second, the number of hidden states for HMM's was chosen to be three, according to the actions of the dataset. Third the best number of clusters for histogram was experimented to be eight, which conforms with the number of classes of states in the proposed method.

Results are listed in Table III. The proposed method using both of the distance methods are compared with the method of Xia et al. [44]. Results reveal that in all actions, the proposed method using each of the two distance measures outperforms the method [44]. Although method [44] has utilized LDA and clustering methods in preparing data for training HMM, it has made several states very close to each other by using a histogram concept, which has increased the error. As an example, in fall actions the angular positions of joints are much similar and the use of histogram ignores their differences.

Using Mahalanobis distance has significantly enhanced the performance, especially in fall actions. In other words, improving the performance of recognizing difficult involuntary states such as crouching and lay front, has improved the total recognition rate. As mentioned before, the main reason of this fact is that the intrinsic variance of states are considered in Mahalanobis distance.

The confusion matrix of actions is reported in Fig. 12. This matrix shows that the actions that are similar to each other are sometimes confused and wrongly recognized. Actions such as falling on front, side and back are sometimes confused with each other because their distribution (and therefore their behavior) are similar and wider than others, as is obvious in Fig. 6. In some scenarios such as anomaly detection in actions, this confusion between subgroup actions might not matter. Hence, another experiment was performed considering all fall and all normal actions as two different high-level groups. In this scenario, the recognition rate improves from 88.64% to 96.18%. And as can be seen in Table IV, false alarm rate has also been significantly reduced. This result indicates that the possibility of wrongly recognizing a normal action as fall action is considerably low.

V. Conclusion and Future Work

A. Conclusion

A new action recognition method was proposed in this paper which is especially useful for recognizing the actions with some sort of complexities such as various types of falling action. Since this method uses feature vectors with low dimension and does not have big computational overhead, it can be used in real time purposes. Experiments showed

that this method outperforms the other methods especially in scenarios where normal and involuntary actions are mixed up.

In the proposed method, a feature vector is created for representing the state of body in each frame, using the Kinect data. The state of body is then recognized in the corresponding discriminative Fisher subspace. Finally, actions are classified and recognized by feeding the sequence of recognized states of body to HMMs. Because of using HMM, this method is robust to different paces and lengths of actions. Moreover, the Mahalanobis distance is utilized for considering the wider distribution of involuntary body states in order to enhance the recognition rate.

B. Potential Future Work

Data was preprocessed by skeleton alignment, to make the algorithm robust against the orientation of camera. As future work, the angles between the joints can be used instead of their positions in order to get more robustness. In addition, recognizing more complex and longer actions can be considered as future work.

Moreover, manual selection/sampling of body states limits the scalability of the system. Automatic selection of body states in an approach similar to [20] which automatically finds elementary states of higher level actions, can also be considered as future work.

Another possible limitation of the proposed method is that canceling the motion of body by alignment, which is necessary for the proposed method, omits the motion information. This cancellation might cause difficulties in recognizing actions with close body states but different motions. Handling this issue can be considered as another potential future work.

VI. Acknowledgment

This work was supported by a grant from Iran National Science Foundation (INSF).

References

[1] Tst fall detection dataset. https://ieee-dataport.org/documents/tst-fall-detection-dataset-v2, Accessed: July 15, 2017.

[2] Aggarwal, J.K. and Xia, L. "Human activity recognition from 3d data: A review", *Pattern Recognition Letters*, **48**, pp. 70-80 (2014).

[3] Althloothi, S., Mahoor, M.H., Zhang, X., and Voyles, R.M. "Human activity recognition using multi-features and multiple kernel learning", *Pattern recognition*, **47**(5), pp. 1800-1812 (2014).

[4] Amor, B.B., Su, J., and Srivastava, A. "Action recognition using rate-invariant analysis of skeletal shape trajectories", *IEEE transactions on pattern analysis and machine intelligence*, **38**(1), pp. 1-13 (2016).

[5] Anjum, M.L., Ahmad, O., Rosa, S., Yin, J., and Bona, B. "Skeleton tracking based complex human activity recognition using kinect camera", In *International Conference on Social Robotics, Springer*, pp. 23-33 (2014).

[6] Bishop, C., "Pattern recognition and machine learning", Springer, New York (2007).

[7] Chen, C., Liu, K., and Kehtarnavaz, N. "Real-time human action recognition based on depth motion maps", *Journal of real-time image processing*, **12**(1), pp. 155-163 (2016).

[8] Chen, C., Liu, M., Zhang, B., Han, J., Jiang, J., and Liu, H. "3d action recognition using multi-temporal depth motion maps and fisher vector", In *IJCAI*, pp. 3331-3337 (2016).

[9] Chen, L., Wei, H., and Ferryman, J. "A survey of human motion analysis using depth imagery", *Pattern Recognition Letters*, **34**(15), pp. 1995-2006 (2013).

[10] Fisher, R.A. "The use of multiple measures in taxonomic problems", *Annals of Eugenics*, **7**, pp. 179-188 (1936).

[11] Gasparrini, S., Cippitelli, E., Gambi, E., Spinsante, S., Wåhslén, J., Orhan, I., and Lindh, T. "Proposal and experimental evaluation of fall detection solution based on wearable and depth data fusion", In *ICT innovations, Springer*, pp. 99-108 (2016).

[12] Guo, K., Ishwar, P., and Konrad, J. "Action recognition from video using feature covariance matrices", *IEEE Transactions on Image Processing*, **22**(6), pp. 2479-2494 (2013).

[13] Han, J., Shao, L., Xu, D., and Shotton, J. "Enhanced computer vision with microsoft kinect sensor: A review", *IEEE transactions on cybernetics*, **43**(5), pp. 1318-1334 (2013).

[14] Hastie, T., Tibshirani, R., and Friedman, J. "The Elements of Statistical Learning: Data Mining, Inference, and Prediction", 2nd Edn., Springer, New York (2009).

[15] Holte, M.B., Tran, C., Trivedi, M.M., and Moeslund, T.B. "Human action recognition using multiple views: a comparative perspective on recent developments", In *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding, ACM*, pp. 47-52 (2011).

[16] Hussein, M.E., Torki, M., Gowayyed, M.A., and El-Saban, M. "Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations", In *IJCAI*, **13**, pp. 2466-2472 (2013).

[17] Huynh, L., Ho, T., Tran, Q., Dinh, T.B., and Dinh, T. "Robust classification of human actions from 3d data", In *Signal Processing and Information Technology (ISSPIT), 2012 IEEE International Symposium on*, pp. 263-268 (2012).

[18] Kapsouras, I. and Nikolaidis, N. "Action recognition on motion capture data using a dynemes and forward differences representation", *Journal of Visual Communication and Image Representation*, **25**(6), pp. 1432-1445 (2014).

[19] Ke, Q., Bennamoun, M., An, S., Sohel, F., and Boussaid, F. "A new representation of skeleton sequences for 3d action recognition", *arXiv preprint arXiv:1703.03492* (2017).

[20] Lee, S., Le, H.X., Ngo, H.Q., Kim, H.I., Han, M., Lee, Y.K., et al. "Semi-markov conditional random fields for accelerometer-based activity recognition", *Applied Intelligence*, **35**(2), pp. 226-241 (2011).

[21] Li, W., Zhang, Z., and Liu, Z. "Action recognition based on a bag of 3d points", In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pp. 9-14 (2010).

[22] Liang, C., Chen, E., Qi, L., and Guan, L. "3d action recognition using depth-based feature and locality-constrained affine subspace coding", In *Multimedia (ISM), 2016 IEEE International Symposium on*, pp. 261-266 (2016).

[23] Liu, A.A., Nie, W.Z., Su, Y.T., Ma, L., Hao, T., and Yang, Z.X. "Coupled hidden conditional random fields for rgb-d human action recognition", *Signal Processing*, **112**, pp. 74-82 (2015).

[24] Liu, J., Luo, J., and Shah, M. "Recognizing realistic actions from videos in the wild", In *Computer vision and pattern recognition, 2009. CVPR 2009. IEEE conference on*, pp. 1996-2003 (2009).

[25] Liu, J., Shahroudy, A., Xu, D., and Wang, G. "Spatio-temporal lstm with trust gates for 3d human action recognition", In *European Conference on Computer Vision, Springer*, pp. 816-833 (2016).

[26] Liu, L, Shao, L., Li, X., and Lu, K. "Learning spatio-temporal representations for action recognition: A genetic programming approach", *IEEE transactions on cybernetics*, **46**(1), pp. 158-170 (2016).

[27] Liu, M., Chen, C., Meng, F., and Liu, H. "3d action recognition using multi-temporal skeleton visualization", In *Multimedia & Expo Workshops (ICMEW), 2017 IEEE International Conference on*, pp. 623-626, (2017).

[28] Luvizon, D.C., Tabia, H., and Picard, D. "Learning features combination for human action recognition from skeleton sequences", *Pattern Recognition Letters* (2017).

[29] Martínez-Zarzuela, M, Díaz-Pernas, FJ, Tejeros-de-Pablos, A, González-Ortega, D, and Antón-Rodríguez, M. "Action recognition system based on human body tracking with depth images", *Advances in Computer Science: an International Journal*, **3**(1), pp. 115-123 (2014).

[30] Niebles, J.C., Wang, H., and Fei-Fei, L. "Unsupervised learning of human action categories using spatial-temporal words", *International journal of computer vision*, **79**(3), pp. 299-318 (2008).

[31] Papadopoulos, G.T., Axenopoulos, A., and Daras, P. "Real-time skeleton-tracking-based human action recognition using kinect data", In *MMM* **1**, pp. 473-483 (2014).

[32] Peng, X., Wang, L., Wang, X., and Qiao, Y. "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice", *Computer Vision and Image Understanding*, **150**, pp. 109-125 (2016).

[33] Rabiner, L.R. "A tutorial on hidden markov models and selected applications in speech recognition", In *Proceedings of the IEEE*, **77**(2), pp. 257-286 (1989).
[34] Rahmani, H., Mahmood, A., Huynh, D., and Mian, "A. Histogram of oriented principal components for cross-view action recognition", *IEEE transactions on pattern analysis and machine intelligence*, **38**(12), pp. 2430-2443 (2016).
[35] Rahmani, H., and Mian, A. "3d action recognition from novel viewpoints", In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1506-1515 (2016).
[36] Reddy, V.R. and Chattopadhyay, T. "Human activity recognition from kinect captured data using stick model", In *International Conference on Human-Computer Interaction, Springer*, pp. 305-315 (2014).
[37] Shahroudy, A., Ng, T.T., Yang, Q., and Wang, G. "Multimodal multipart learning for action recognition in depth videos", *IEEE transactions on pattern analysis and machine intelligence*, **38**(10), pp. 2123-2129 (2016).
[38] Theodorakopoulos, I., Kastaniotis, D., Economou, G., and Fotopoulos, S. "Pose-based human action recognition via sparse representation in dissimilarity space", *Journal of Visual Communication and Image Representation*, **25**(1), pp. 12-23 (2014).
[39] Theodoridis, S. and Koutroumbas, K. "Pattern recognition", 2nd Edn., Elsevier Academic Press, USA (2003).
[40] Wang, H., Kläser, A., Schmid, C., and Liu, C.L. "Action recognition by dense trajectories", In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 3169-3176 (2011).
[41] Wang, H., Kläser, A., Schmid, C., and Liu, C.L. "Dense trajectories and motion boundary descriptors for action recognition", *International journal of computer vision*, **103**(1), pp. 60-79 (2013).
[42] Wang, H., Oneata, D., Verbeek, J., and Schmid, C. "A robust and efficient video representation for action recognition", *International Journal of Computer Vision*, **119**(3), pp. 219-238 (2016).
[43] Wang, J., Liu, Z., and Wu, Y. "Learning actionlet ensemble for 3d human action recognition", In *Human Action Recognition with Depth Cameras, Springer*, pp. 11-40 (2014).
[44] Xia, L., Chen, C.C., and Aggarwal, J.K. "View invariant human action recognition using histograms of 3d joints", In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pp. 20-27 (2012).
[45] Yao, A., Gall, J., and Gool, L.V. "Coupled action recognition and pose estimation from multiple views", *International journal of computer vision*, **100**(1), pp. 16-37 (2012).
[46] Ye, M., Zhang, Q., Wang, L., Zhu, J., Yang, R., and Gall, J. "A survey on human motion analysis from depth data", In *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications, Springer*, pp. 149-187 (2013).
[47] Zhang, B., Yang, Y., Chen, C., Yang, L., Han, J., and Shao, L. "Action recognition using 3d histograms of texture and a multi-class boosting classifier", *IEEE Transactions on Image Processing*, **26**(10), pp. 4648-4660 (2017).
[48] Zhao, W., Chellappa, R., Phillips, P.J., and Rosenfeld, A. "Face recognition: A literature survey", *ACM computing surveys (CSUR)*, **35**(4), pp. 399-458 (2003).
[49] Zhao, Y., Liu, Z., Yang, L., and Cheng, H. "Combing rgb and depth map features for human activity recognition", In *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific, IEEE*, pp. 1-4 (2012).

Mozhgan Mokari received her BSc degree in electrical engineering from Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran, in 2014. She also received her MSc degree in electrical engineering from Sharif University of Technology, Tehran, Iran, in 2016. She is currently studying for PhD of electrical engineering in Sharif University of Technology. Her research interests are machine learning, computer vision and signal processing.

Hoda Mohammadzade received her BSc degree from Amirkabir University of Technology (Tehran Polytechnic), Iran, in 2004, the MSc degree from the University of Calgary, Canada, in 2007, and the PhD degree from the University of Toronto, Canada, in 2012, all in electrical engineering. She is currently an assistant professor of electrical engineering at Sharif University of Technology, Tehran, Iran. Her research interests include signal and image processing, computer vision, pattern recognition, biometric systems, and bioinformatics.

Benyamin Ghojogh obtained his first and second BSc degrees in electrical engineering (electronics and telecommunications) from Amirkabir University of technology, Tehran, Iran, in 2015 and 2017 respectively. He also received his MSc degree in electrical engineering (digital electronic systems) from Sharif University of technology, Tehran, Iran, in 2017. He is currently studying for PhD of electrical and computer engineering in University of Waterloo, Canada. One of his honors is taking the second rank of Electrical Engineering Olympiad of Iran in 2015. His research interests include machine learning and computer vision.

Captions of figures:

- Figure 1: Applications of human action recognition.
- Figure 2: A human action recognition system for fall detection.
- Figure 3: Selected joints out of available joints in the skeletal data. The joints used for alignment are also shown.
- Figure 4: Alignment of skeleton using the left and right shoulders to cancel the orientation of skeleton.
- Figure 5: An example of Fisher and PCA directions.
- Figure 6: Projection of samples of states onto Fisher space. As can be seen, the states have different distributions.
- Figure 7: A three-state HMM model.
- Figure 8: The overall structure of proposed framework.
- Figure 9: An example of actions in TST dataset [1].
  - Figure 9-a: Sit
  - Figure 9-b: Grasp
  - Figure 9-c: Walk
  - Figure 9-d: Lay
  - Figure 9-e: Fall front
  - Figure 9-f: Fall back
  - Figure 9-g: Fall side
  - Figure 9-h: End up sit
- Figure 10: An example of the selected states.
  - Standing
  - Crouching
  - Lay back
  - Lay front
  - Lay side
  - Bend
  - Sit on chair
  - Sit on ground
- Figure 11: Confusion matrix of states.
  - Figure 11-a: Euclidean
  - Figure 11-b: Mahalanobis
- Figure 12: Confusion matrix of actions.
  - Figure 11-a: Euclidean
  - Figure 11-b: Mahalanobis

Captions of tables:

- Table 1: Correctness rate of recognizing state of body
- Table 2: Effect of number of states of HMM on the recognition rate
- Table 3: Comparison of results of our method and method [44] for TST dataset
- Table 4: Comparison of results, considering all abnormal actions to be fall event

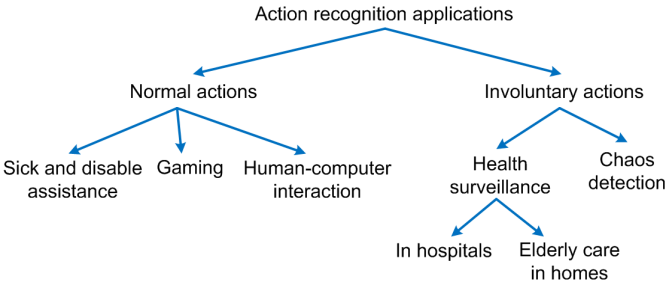


Fig. 1: Applications of human action recognition.

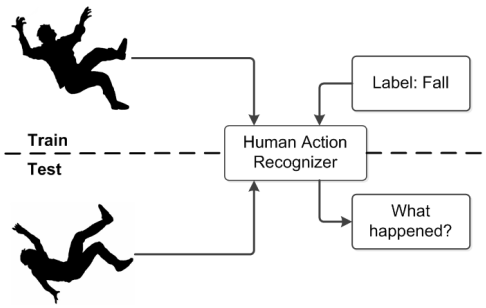


Fig. 2: A human action recognition system for fall detection.

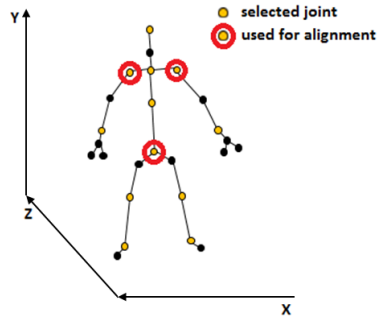


Fig. 3: Selected joints out of available joints in the skeletal data. The joints used for alignment are also shown.

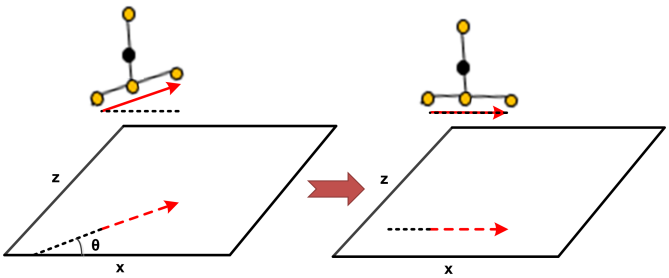
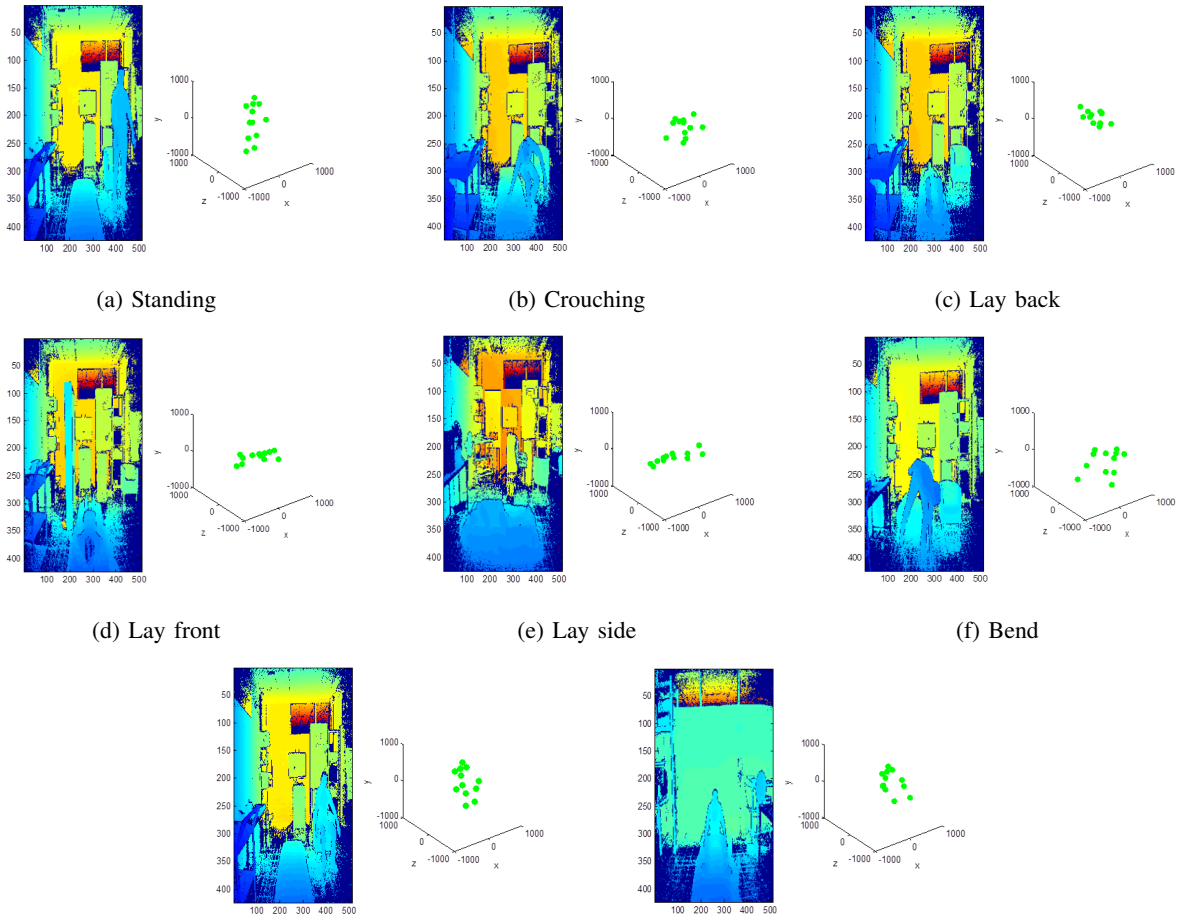


Fig. 4: Alignment of skeleton using the left and right shoulders to cancel the orientation of skeleton.



(a) Standing (b) Crouching (c) Lay back

(d) Lay front (e) Lay side (f) Bend

(g) Sit on chair (h) Sit on ground

Fig. 10: An example of the selected states.

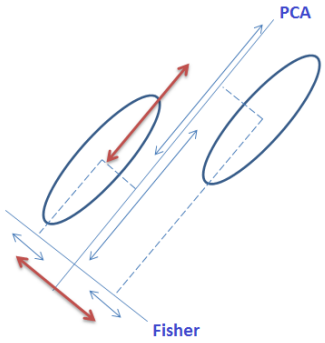


Fig. 5: An example of Fisher and PCA directions.

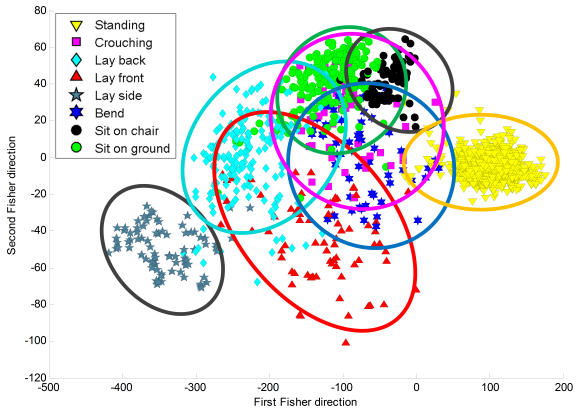


Fig. 6: Projection of samples of states onto Fisher space. As can be seen, the states have different distributions.

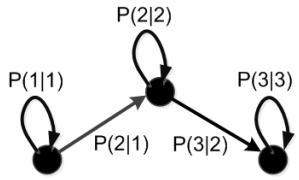


Fig. 7: A three-state HMM model.

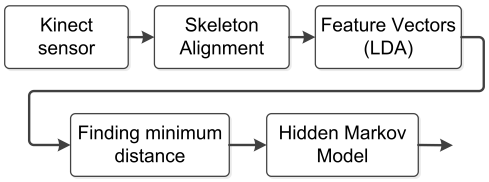
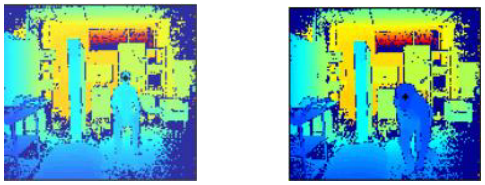


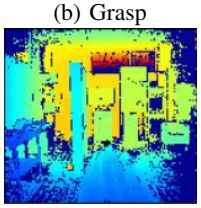
Fig. 8: The overall structure of proposed framework.



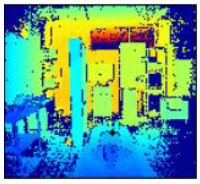
(a) Sit

(b) Grasp

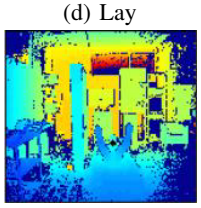
(c) Walk



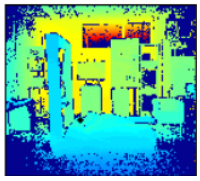
(d) Lay



(e) Fall front



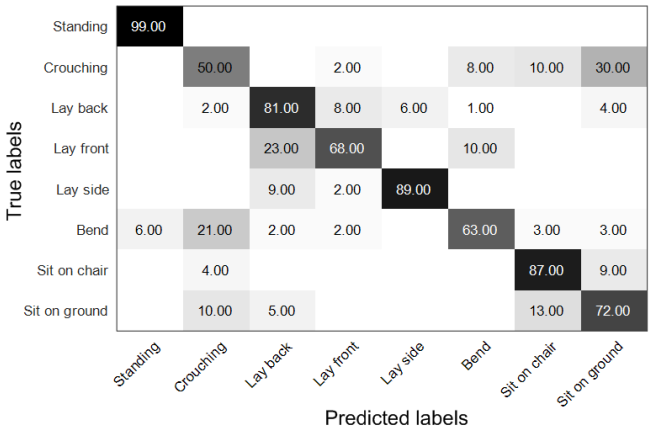
(f) Fall back



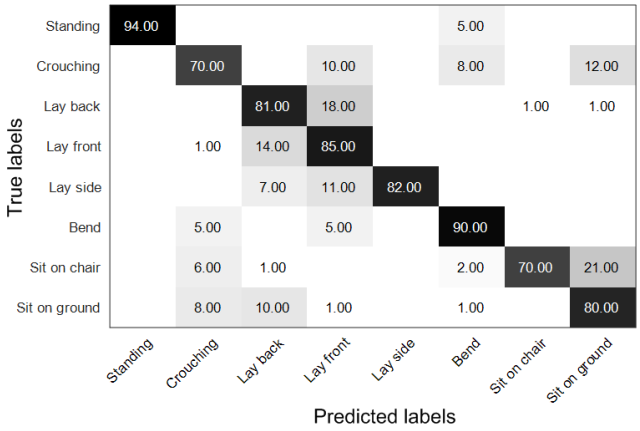
(g) Fall side

(h) End up sit

Fig. 9: An example of actions in TST dataset [1].



(a) Euclidean

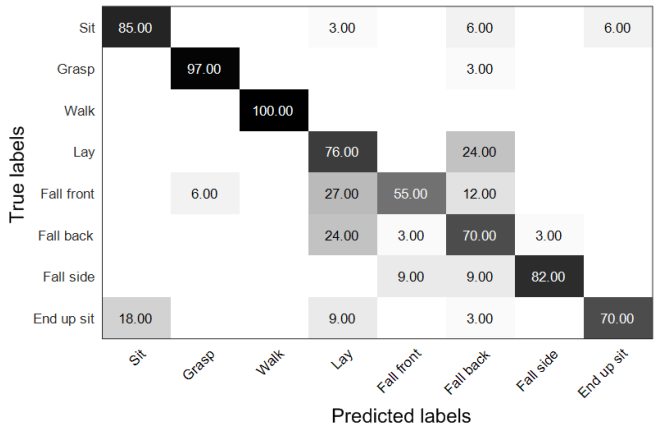


(b) Mahalanobis

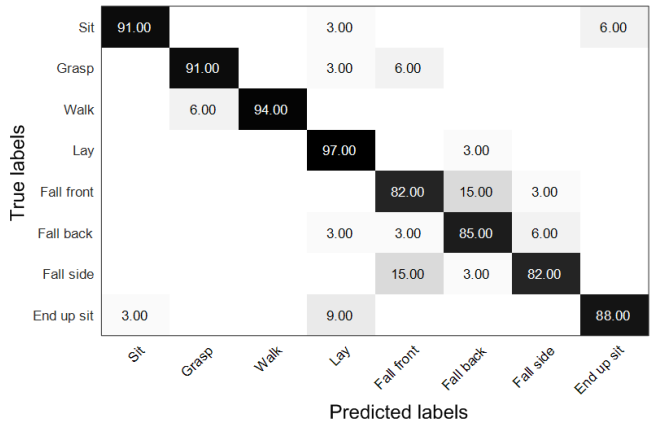
Fig. 11: Confusion matrix of states.

TABLE I: Correctness rate of recognizing state of body

State	Euclidean	Mahalanobis
Standing	99.38%	94.26%
Crouching	50.00%	70.00%
Lay back	80.71%	81.22%
Lay front	67.50%	85.00%
Lay side	88.89%	82.22%
Bend	62.90%	90.32%
Sit on chair	86.87%	69.70%
Sit on ground	72.15%	79.91%
Total	**76.03%**	**81.57%**



(a) Euclidean



(b) Mahalanobis

Fig. 12: Confusion matrix of actions.

TABLE III: Comparison of results of our method and method [44] for TST dataset

Action	Euclidean	Mahalanobis	[44]
Sit	84.85%	90.91%	81.82%
Grasp	96.97%	90.91%	84.85%
Walk	100%	93.94%	90.91%
Lay	75.76%	96.97%	90.91%
Fall front	54.54%	81.82%	48.49%
Fall back	69.70%	84.85%	66.67%
Fall side	81.82%	81.82%	69.70%
End up sit	69.70%	87.88%	33.33%
Total	79.16%	**88.64%**	70.83%

TABLE II: Effect of number of states of HMM on the recognition rate

Action	2 states	3 states	4 states
Sit	87.88%	90.91%	90.91%
Grasp	90.91%	90.91%	87.88%
Walk	93.94%	93.94%	93.94%
Lay	84.85%	96.97%	90.91%
Fall front	84.85%	81.82%	81.82%
Fall back	84.85%	84.85%	78.79%
Fall side	81.82%	81.82%	81.82%
End up sit	84.85%	87.88%	84.85%
Total	86.74%	**88.64%**	86.36%

TABLE IV: Comparison of results, considering all abnormal actions to be fall event

	Euclidean	Mahalanobis	[44]
Recognition Rate (true positive rate)	78.78%	**96.18%**	77.27%
Specificity Rate (true negative rate)	90.15%	**96.21%**	90.90%
False Alarm Rate (false positive rate)	9.15%	**3.78%**	9.09%