Human Activity Recognition Based on Weighted Limb Features

Liang Zhang, Wenhan Yang, Guangming Zhu, Peiyi Shen, and Juan Song

Abstract— Human activity recognition plays an important role in personal assistive robot, being able to recognize human activity and perform corresponding assistive action is a great challenge for personal assistive robot. Human body is an articulated system of rigid segments that can be divided into five parts, but many existing methods always identify actions based on the motion trajectories of whole body. In this paper, taking into account the fact that most actions can be performed by a few limbs and the other limbs should not impact on the action recognition, we proposed an activity recognition method based on limb weights. The weight of each limb is composed of consistency weight and uniqueness weight, which are learned according to the similarity degree among different sequences for each specific action. The covariance descriptor, which is the concatenation of eigenvalues extracted from covariance matrices, is adopted to represent the motion trajectory of each limb. In order to distinguish action instances from each other in the feature sequences, a simple annotation method is used. Experimental results on the Cornell activity dataset and the Lab dataset show that the proposed method not only can outperform the state-of-the-art algorithms, but also is appropriate to recognize the actions whose non-core limbs’ trajectories are different from each other.

I. INTRODUCTION

Recently, the research on robot becomes increasingly hot widely concerned by institutes, colleges and scholars of countries. Being able to recognize human activity and perform corresponding assistive action is a great challenge for personal assistive robots [3], which have huge application prospects in many fields (e.g. smart home, health care, human-computer interaction). There are many existing methods that recognize human activity based on RGB information. However, due to the similarity of inter-class among some different actions and the dissimilarity of intra-class when different persons perform the same action, it is difficult for robots to identify actions robustly only with RGB information. Fortunately, with the release of several low-cost and relatively accurate RGB-D sensors (e.g., the Microsoft Kinect sensor), the collection of RGB-D sequences becomes less difficult. On the other hand, the human body is an articulated system of rigid segments that are connected by joints in 3D space, and human motion can be considered as a temporal evolution of the joints [4]. Microsoft and PrimeSense presented a framework for the extraction of human skeletal data in succession, and then a large number of skeleton-based action recognition methods emerge.

All Authors are with School of Software, Xidian University, Xi’an, China (corresponding author is Guangming Zhu; e-mail: gmzhu@xidian.edu.cn).
This work is partially supported by the Natural Science Foundation of China Grant (NO. 61401324, NO. 61305109, NO. 61072105), by 863 Program (2013AA014601), and by Shaanxi Scientific Research Plan (2014K07-11).

Fig. 1 Overview of the proposed human activity recognition algorithm.

Sung et al. [2] extract a set of features from RGB and skeletal data, which can effectively represent human pose and motion, and then put these features in the hierarchical maximum entropy Markov model to perform recognition. Sempena et al. [5] build feature vector from joint orientation along time series, and apply dynamic time warping algorithm to these feature vectors. A new body part-based skeleton representation is proposed to describe the rotation and translation relation between body parts, and human actions are modeled as curves in a Lie group [6]. Xia et al. [7] present a new view-invariant posture representation by computing histogram of joint positions which are mapped to a spherical coordinate system.

Most methods described above are based on the information of whole body, but many actions can be accurately recognized by identifying the trajectory of a few limbs. For example, the action “drinking water” can be recognized based on the motion information of upper body regardless of the state of lower body (e.g., people are standing, sitting or walking). However, most of the existing methods identify activities based on the posture change of whole body. This will lead to low precision when some limbs’ trajectories of the actions are different. In order to resolve the problem, a new representation of human actions called sequence of the most informative of joints (SMIJ) is proposed in [8], which selects a few most informative skeleton joints and represents an action as a feature sequence of these joints instead of whole body. But the method has some problems: Firstly, the most informative joints sequences of different actions are not exactly the same, and there is no comparison among the features computed over these sequences. Secondly, some of the chosen joints are not at the core limbs for certain actions, which will disturb the recognition result.

The human body can be divided into five parts (i.e. torso, left arm, right arm, left leg and right leg). Generally, it is
enough to perform most actions using a few limbs, thus several actions can be performed at the same time (e.g., calling and walking). Therefore, we can recognize each action based on the most informative limbs instead of whole body. According to this idea, we propose to divide the feature sequence into five parts based on five limbs. In order to avoid the above situation that the feature sequences extracted from different limbs cannot be compared, we introduce limb weight to show the importance of different limbs to each activity.

In this paper, we propose a new method which is easy to learn limb weights based on specific action characteristics for each sequence. Firstly, all joints are transformed into local coordinate system with respect to the torso, and the position sequence is divided into five parts. Besides, the preprocessing operation on skeletal data could overcome the tremble phenomenon of human poses. Secondly, several covariance matrices are calculated over each action sequence, and N eigenvalues are extracted from a covariance matrix of order N, then the concatenation of these eigenvalues is regarded as the covariance descriptor which can represent the motion trajectory of a limb in a specific time. Thirdly, the limb weight is learned according to the similarity degree among different sequences for each action, which is divided into two parts: the consistency weight and the uniqueness weight. Lastly, the limb weight is combined with the classification method using key poses and atomic motions [1]. An overview of the proposed method is shown in Figure 1.

The remainder of the paper is structured as follows. Section II describes the learning process of limb weight. The action classification method is introduced in Section III. Section IV gives the experimental results and makes an analysis. Finally, Section V summarizes the paper and discusses the future directions.

II. LEARNING WEIGHTS OF LIMBS

This section gives a detailed description of the learning process of limb weights. In our proposed method, the weight is divided into the consistency weight and the uniqueness weight which respectively represent the similarity of several motion trajectories in the sequences that are of the same category for each limb, and the dissimilarity of several motion trajectories in the sequences that are of the different category for each limb. In the following, we first give the preprocessing and normalization of the skeletal data, then the covariance descriptors is described, and finally the consistency weight and the uniqueness weight are illustrated.

A. Preprocessing and normalization

Human body is an articulated system of rigid segments which can be represented by several joints, these joints in 3D space can be easily obtained from RGB-D sensors. But some extracted skeletal data are unstable and noisy due to the limitation of the accuracy of sensors and the complexity of detection environment that makes observed human poses tremble and has an adverse impact on the recognition accuracy. It is necessary to preprocess the skeletal data by utilizing a moving average on them. The filter is given by the equation

\[ f'_n = \frac{1}{w} \sum_{w=0} f_{n+w-w/2}, \]  

\[ (1) \]

where \( f_n \) is the value of the \( n \)th pre-filtered data and \( f'_n \) is the filtered one, \( W \) is the length of the moving window.

A fact is that different persons’ height, limb length, distance to RGB-D sensor and the orientation they face in different scenarios are different. Because the original coordinates of skeletal joints obtained from sensors are sensitive to the detection environment, normalization operation should be performed on the skeletal data, and in our implementation the relative position of each joint with respect to the torso is calculated as

\[ \overline{P}_{tx} = R_t \frac{P_x - P_t}{\|P_x - P_t\|}, \]  

\[ (2) \]

where \( P_t \) is the torso and \( P_x \) is the one of other joints. \( R_t \) and \( \|P_x - P_t\| \) are respectively the rotation and the Euclidean distance between the torso and joint \( x \).

B. The covariance descriptor

1) Extraction of multiple covariance matrices

The change process of human pose can be denoted as five limbs’ trajectories in an action sequence. In our proposed method, we use the covariance matrix for normalized skeletal joint locations over time as a discriminative descriptor for a trajectory. In this case, it can be easily known that the length of covariance descriptor is fixed, and is independent from the length of the action sequence.

Suppose that a limb consists of \( K \) joints, and the action is performed over \( T \) frames. Let \( S(t) \) be the vector of \( K \) joint locations, that is \( S(t) = [x_1(t), y_1(t), z_1(t), \ldots, x_K(t), y_K(t), z_K(t)] \), where \((x_i(t), y_i(t), z_i(t))\) is the coordinate of the \( i \)th joint at frame \( t \). And let \( S \) be the collection of these vectors, that is \( S = [S(1); S(2); \ldots; S(T)] \). Then, the covariance matrix \( C(S) \) can be calculated as

\[ C(S) = \frac{1}{T-1} (S - \overline{S})(S - \overline{S}), \]  

\[ (3) \]

where \( \overline{S} \) is the sample mean of \( S \).

Taking into consideration the fact that covariance matrix cannot capture the order of trajectory in time, temporal information will be lost if the trajectory is just represented by one covariance matrix. Worse, when two actions are the reverse temporal order of one another, the two calculated covariance matrices will be same completely. In order to emphasize the time relationship in the sequence, we use multiple covariance matrices over sub-sequence, which is
presented in [9]. In our approach, the lower level matrices are computed over smaller sub-sequences, which include \( T/2^{h-1} \) frames, where \( h \) is the level of the matrix. The step from one sub-sequence to the next is half length of the sub-sequence. Figure 2 shows the hierarchical structure of multiple covariance matrices.

2) Eigenvalues of covariance matrix

From the definition of the covariance matrix, it can be easily seen that it is a symmetric \( N \times N \) matrix, which can describe the correlation of different joints with one another. However, it is unrealistic for all sample data to distribute along the coordinate axes in a fixed coordinate system. Therefore, the upper triangle of the covariance matrix cannot be directly utilized as the descriptor due to the influence of the data distribution in original dimensional space on the covariance matrix. Given the fact that a symmetric matrix of order \( N \) can be decomposed into \( N \) Eigenvectors according to the distribution of the corresponding sample data, and the matrix will be mapped to the \( N \) orthonormal basis. Eigenvectors can be obtained by solving the equation

\[
(A - \lambda I)x = 0,
\]

where \( A \) is the covariance matrix, \( \lambda \) is the eigenvector and \( x \) is the eigenvalue which represents the projection length of the matrix on this eigenvector.

Due to the direct proportional relation between eigenvalue and the variance of the sample data on the corresponding eigenvector, eigenvalue can effectively represent the amount of information in eigenvector. Thus the set of \( N \) eigenvalues can discriminatively describe the corresponding covariance matrix.

As mentioned above, multiple covariance matrices are computed over the entire sequence or its sub-sequences, which can represent the time relationship in the sequence. Therefore, the covariance descriptor can be calculated as

\[
F = [X^{(1)}, X^{(2)}, \ldots, X^{(M)}]^T,
\]

which describes the trajectory of a limb in a specific time, where \( X^{(m)} \) is the vector of \( N \) eigenvalues extracted from the \( m \)th covariance matrix, i.e. \( X^{(m)} = [x_1^{(m)}, x_2^{(m)}, \ldots, x_N^{(m)}] \), \( M \) is the number of covariance matrices, then one covariance descriptor has \( E = M \times N \) elements.

C. Consistency weight

As aforementioned, human body is composed of five limbs which are respectively torso, left arm, right arm, left leg and right leg. It can be easily recognized that which action is performed during the sequence by identifying the trajectories of these five parts. When all motion trajectories of a limb (e.g. left arm) belong to the same action are similar, the limb will be regarded as the core part. On the contrary, the limb will be considered as the non-core part for an action if a part of the limb’s motion trajectories are far different from the others in the sequences which are of the same category.

According to the computation process of the covariance descriptor in the last subsection, the similarity problem of the motion trajectories can be converted into the distance problem between the corresponding descriptors. But the magnitude of a covariance descriptor which represents the amplitude of the motion trajectory is useless to estimate the type of each trajectory. It is more reasonable to utilize cosine similarity to describe the angles among different covariance descriptors. Let \( \bar{F}^{(c)}_l \) be the mean of \( D \) descriptors belong to the \( l \)th limb for the \( c \)th action, that is \( \bar{F}^{(c)}_l = \frac{1}{D} \sum_{d=1}^{D} F^{(c)}_{d,l} \), then the angle between the \( d \)th covariance descriptor \( F^{(c)}_{d,l} \) and \( \bar{F}^{(c)}_l \) can be calculated as

\[
\theta_{d,l}^{(c)} = \arccos \left( \frac{\bar{F}^{(c)}_{d,l} \bar{F}^{(c)}_l}{|\bar{F}^{(c)}_{d,l}| \bar{F}^{(c)}_l} \right),
\]

where \(|\cdot|\) is the length of a vector, then the mean angle for a limb can be calculated as

\[
\bar{\theta}_l^{(c)} = \frac{1}{D} \sum_{d=1}^{D} \theta_{d,l}^{(c)}.
\]

The smaller the mean angle among different descriptors is, the more similar the corresponding trajectories will be. Thus the consistency weight can be computed as

\[
\text{ConW}_l^{(c)} = \frac{1}{\bar{\theta}_l^{(c)}}.
\]

D. Uniqueness weight

In contrast with the consistency weight, the uniqueness weight is used to represent the dissimilarity of the motion trajectories of a limb in the sequences which are of the different category. When the limb’s motion trajectories of an action are far different from other actions’ trajectories, the trajectory of the limb can effectively represent this action. More specifically, if the trajectory of a limb appears only in a few action sequences, the limb should be given larger uniqueness weight.

Given the mean of \( D \) covariance descriptors belong to the \( l \)th limb for the \( c \)th action is denote as \( \bar{F}^{(c)}_l \), the angle between mean descriptor \( c \) and \( q \) can be calculated as

\[
\phi_l^{(c,q)} = \arccos \left( \frac{\bar{F}^{(c)}_l \bar{F}^{(q)}_l}{|\bar{F}^{(c)}_l| \bar{F}^{(q)}_l} \right).
\]

Let \( \phi_l^{(c)} \) be the sum of the angles among mean descriptor \( c \) and the other descriptors, that is \( \phi_l^{(c)} = \sum_{q=1}^{C} \phi_l^{(c,q)} \), where \( C \) is the number of activity. Then the uniqueness weight can be computed as

\[
\text{UniW}_l^{(c)} = \frac{\phi_l^{(c)}}{\sum_{e=1}^{C} \phi_l^{(e)}}.
\]

III. ACTION CLASSIFICATION METHOD

In the action classification process, we utilize an effective and efficient action classification method proposed in [1], which is combined with the limb weight learned in the last section. Due to the fact that one human action can be considered as a sequence of key poses and atomic motions, in the reference method, it first segments feature sequences into static segments and dynamic segments based on the kinetic energy, and then the clustering algorithms are used to extract key poses and atomic motions from static and dynamic
segments respectively. After that, an action pattern like “key pose - atomic motion - key pose” is proposed and the length of the action pattern is set to an empirical value. Lastly, the Naïve Bayes Nearest Neighbor (NBNN) algorithm is used to classify human actions based on the action pattern. The classification result can be computed as

\[ C^* = \arg\min_{c} \sum_{l=1}^{L} \left( \text{ConW}_l^{(c)} \cdot \text{UnW}_l^{(c)} \cdot \sum_{p=1}^{P} \text{Dist}_{l,p}^{(c)} \right), \]

(11)

where \( C^* \) is the type estimated by classification model over the testing sequence, \( L \) and \( P \) are the number of limbs and components of the pattern respectively. \( \text{Dist}_{l,p}^{(c)} \) is the Euclidean distance between the \( p^{th} \) component of the testing pattern and the best-match training pattern of the \( c^{th} \) activity for the \( l^{th} \) limb. The Euclidean distance of “key pose” components can be calculated as

\[ \text{Dist}_{l,p}^{(c)} = \| F_l^{(c)} - K_{p_{l,c}}^{\text{best}} \|^2, \]

(12)

where \( F_l^{(c)} \) and \( K_{p_{l,c}}^{\text{best}} \) are respectively the testing feature vector and the best-match key pose of the \( c^{th} \) activity for the \( l^{th} \) limb, and the Euclidean distance of “atomic motion” components can be calculated as

\[ \text{Dist}_{l,p}^{(c)} = \| A_{l}^{(c)} - A_{l_{p,c}}^{\text{best}} \|^2, \]

(13)

where \( A_{l}^{(c)} \) and \( A_{l_{p,c}}^{\text{best}} \) are respectively the atomic motion of testing pattern and the best-match atomic motion of the \( c^{th} \) activity for the \( l^{th} \) limb.

IV. EXPERIMENT RESULTS AND ANALYSIS

In this section, we evaluate the proposed method with the state-of-the-art methods on the Cornell activity dataset and employ the same experimental setup to demonstrate the superiority of the proposed method. The comparative results show that the proposed method can outperform the performance of state-of-the-art algorithms.

In addition, take into consideration the fact that several actions are performed at the same time in daily life, the recognition result of an action will be interfered by other actions performed by non-core limbs. Thus a new activity dataset called Lab activity dataset is collected and several actions are performed synchronously in each sequence. Then, we evaluate the proposed method with the state-of-the-art algorithms on the Lab activity dataset. We find that the proposed method outperforms the other methods on the Lab activity dataset, and this can demonstrate the effectiveness of limb weights. It is worth noting that all experiments are performed using three levels in the hierarchy.

A. Cornell activity dataset

Cornell activity dataset (CAD-60) consists of the skeleton data, RGB and depth sequences obtained by the Microsoft Kinect sensor. Twelve high-level actions are performed by four human subjects: two males and two females, which can be categorized into five different environments: bathroom, bedroom, kitchen, living room, office. Three to four common actions are identified for each environment. The twelve kinds of actions are described as follows: brushing teeth, cooking (chopping), cooking (stirring), drinking water, opening pill container, relaxing on couch, rinsing mouth with water, talking on couch, talking on the phone, wearing contact lenses, working on computer, writing on white board.

In this paper, we employ the same experimental setup described in [2]. Leave-one-out cross validation is utilized in the “new person” setting, i.e. the model is trained on three of the four subjects’ data, and tested on the left one. That the classification is made over unseen testing data ensures the credibility of the recognition results. Thus all experiments are performed with the more challenging “new person” setting.

However, there is a problem with our method that the covariance descriptors are computed over the action instances instead of the sequences including numerous instances. Thus these action instances in each sequence should be annotated before limb weight learning. We observe that there is a standard pose between two adjacent action instances in CAD-60, i.e. the people will return to the standard pose after the completion of a movement, then perform the next action. Therefore, the action sequences can be segmented into numerous instances by identifying these standard poses.

The precision and recall rates of the proposed method in each location on Cornell activity dataset are shown in TABLE I. The proposed method obtains excellent precision/recall rates of 95.5%/95.2% in the “new person” setting. It is worth noting that the recognition results of the proposed method in the kitchen is also as good as the results in other location, which cannot be achieved in other state-of-the-art methods due to the strong similarity of action “cooking (chopping)” with action “cooking (stirring)” in the kitchen. Figure 3 gives the corresponding confusion matrix between the actions in the “new person” setting. It can be easily found that the proposed method can obtain relatively accurate recognition results of all actions but action “talking on the phone” which is partly recognized as action “drinking water”. This is because the motion trajectory-based method cannot accurately distinguish one action from the other which has similar motion trajectories due to the limitation of the accuracy of RGB-D sensors.

The comparison of the performance among the proposed method and the state-of-the art methods on the Cornell activity dataset is shown in TABLE II. It can be intuitively found that the proposed method outperform state-of-the-art performance in the “new person” setting. But we find that only one activity is performed in each sequence in this dataset by the non-core limbs. However, several actions can be performed at the same time in daily life, and the recognition result of a specific action

<table>
<thead>
<tr>
<th>Location</th>
<th>Prec. (%)</th>
<th>Rec. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bathroom</td>
<td>96.4</td>
<td>97.1</td>
</tr>
<tr>
<td>Bedroom</td>
<td>94.6</td>
<td>94.1</td>
</tr>
<tr>
<td>Kitchen</td>
<td>96.7</td>
<td>96.7</td>
</tr>
<tr>
<td>Living room</td>
<td>95.4</td>
<td>94.7</td>
</tr>
<tr>
<td>Office</td>
<td>94.3</td>
<td>93.4</td>
</tr>
<tr>
<td>Average</td>
<td>95.5</td>
<td>95.2</td>
</tr>
</tbody>
</table>
will be interfered by other actions performed by non-core limbs. Therefore, we collect a new activity dataset to demonstrate that the proposed method can solve this problem. All trajectories of the non-core limbs are different from each other for a specific action in the dataset. A detailed description of the dataset is given in the following subsection.

B. Lab activity dataset

Lab activity dataset contains 112 action instances collected by Openni2/NITE framework. Each frame in these instances is composed of fifteen joints which are divided into five parts as aforementioned. All action instances are categorized into two environments: indoor and outdoor, and six actions are performed by four subjects for each environment. For each subject, we collect two to three action instances of each action and the trajectory of each instance’s non-core limbs is different from the other one. The set of actions consists of drinking water, throwing, greeting, clapping, jogging, crossing hands, stretching, serving, swinging, hitting the ball, catching the ball, and shooting. Figure 4 lists the four activities of the dataset, three action instances are performed for each activity, and the non-core limbs’ motion trajectory of each instance is different from each other.

We evaluate the proposed method on Lab activity dataset in the “new person” setting with the state-of-the-art methods. The comparison of the performance among these methods is shown in TABLE III. It can be intuitively noticed that the precision and recall rates in the outdoor location is better than the recognition results in the indoor location for all methods. That is because each outdoor action is far different from the others, but the indoor actions (e.g. “throwing” and “greeting”, “clapping” and “crossing hands”) are similar due to the limitation of the accuracy of RGB-D sensors. However, in the two different environments, the proposed method outperforms other methods. It is noteworthy that the proposed method is based on the statistical characteristics, which is not suitable for the dataset that contains a small amount of data.

<table>
<thead>
<tr>
<th>Method</th>
<th>“New Person” Prec. (%)</th>
<th>“New Person” Rec. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMEMM [2]</td>
<td>67.9</td>
<td>55.5</td>
</tr>
<tr>
<td>HMM + GMM [12]</td>
<td>70.0</td>
<td>78.0</td>
</tr>
<tr>
<td>Eigenjoints [13]</td>
<td>71.9</td>
<td>66.6</td>
</tr>
<tr>
<td>Multilevel Depth and Image Fusion [14]</td>
<td>75.9</td>
<td>69.5</td>
</tr>
<tr>
<td>Depth and Spatial Data Segment [15]</td>
<td>78.1</td>
<td>75.4</td>
</tr>
<tr>
<td>MRF + SSVM [16]</td>
<td>80.8</td>
<td>71.4</td>
</tr>
<tr>
<td>Structure-Motion Feature [17]</td>
<td>86.0</td>
<td>84.0</td>
</tr>
<tr>
<td>DBMM [18]</td>
<td>91.1</td>
<td>91.9</td>
</tr>
<tr>
<td>Self-organizing Neural Integra. [19]</td>
<td>91.9</td>
<td>90.2</td>
</tr>
<tr>
<td>Spatiotemporal Features Fusion [20]</td>
<td>93.2</td>
<td>84.6</td>
</tr>
<tr>
<td>Key Pose and Atomic Motion [1]</td>
<td>94.6</td>
<td>94.8</td>
</tr>
<tr>
<td>Our method</td>
<td>95.5</td>
<td>95.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Indoor Prec. (%)</th>
<th>Indoor Rec. (%)</th>
<th>Outdoor Prec. (%)</th>
<th>Outdoor Rec. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMEMM [2]</td>
<td>65.9</td>
<td>61.1</td>
<td>78.7</td>
<td>73.7</td>
</tr>
<tr>
<td>Key Pose and Atomic Motion [1]</td>
<td>78.5</td>
<td>58.2</td>
<td>95.8</td>
<td>91.3</td>
</tr>
<tr>
<td>Proposed method</td>
<td>84.7</td>
<td>78.0</td>
<td>97.6</td>
<td>95.6</td>
</tr>
</tbody>
</table>
Fig. 4 Four kinds of actions in Lab activity dataset, each action contains three instances and the non-core limbs’ trajectory of each instance is different from each other.

V. CONCLUSION

In this paper, we present a new method to learn limb weights which contributes to the activity recognition. The preprocessing operation on skeletal data can overcome the tremble phenomenon of human poses, and features based on the normalized data are insensitive to persons’ height, limb length, distance and orientation. The covariance descriptor is the concatenation of multiple eigenvalues extracted from several covariance matrices, which can describe the motion trajectory of each limb. The consistency weight and uniqueness weight are calculated using the covariance descriptors respectively. In our experiment, we have shown that the proposed method not only can outperform the state-of-the-art algorithms, but also is appropriate to recognize the actions whose non-core limbs’ trajectories are different from each other. However, the proposed method is only suitable to the sequence that has been annotated with different action instances. In future, we will do more effort to resolve the online automatic instance segmentation for action sequence recognition, and also apply our proposed algorithms to ROS for real robot, such as NAO, recognizing people’s daily activities.

REFERENCES


