Combining 3D Joints Moving Trend and Geometry Property for Human Action Recognition

Bangli Liu[1], Hui Yu[2], Xiaolong Zhou[1,3], Honghai Liu[1]
[1]*School of Computing, University of Portsmouth, UK*
[2]*School of Creative Technologies, University of Portsmouth, UK*
[3]*College of Computer Science and Technology, Zhejiang University of Technology, China*

***Abstract*—Depth image based human action recognition has attracted many attentions due to the popularity of the depth sensors. However, accurate recognition still remains a challenge because of various object appearances, poses and video sequences. In this paper, a novel skeleton joints descriptor based on 3D Moving Trend and Geometry (3DMTG) property is proposed for human action recognition. Specifically, a histogram of 3D moving directions between consecutive frames for each joint is constructed to represent the 3D moving trend feature in spatial domain. The geometry information of joints in each frame is modelled by the relative motion with the initial status. The proposed feature descriptor is evaluated on two popular datasets. The experimental results demonstrate the superior performance of our method over the state-of-the-art methods, especially the higher recognition rates for complex actions.**

***Index Terms*—Human action recognition, 3D Moving Trend, geometry property.**

I. Introduction

As immense applications in human-machine interaction, vedio surveillance, elderly care and entertainment, human action recognition has been attracting extensive attentions in computer vision. Early proposed strategies mainly recognize human action from 2D sequences captured by RGB cameras [1][2][3][4]. However, the sensitivity to illumination changes and subject texture variations often degrades the recognition accuracy. These problems can be solved by using depth information acquired by depth sensors such as Microsoft Kinect and ASUS Xtion, which have been promoting the research on human action recognition. Because images from depth channel provide another dimension information (the depth data), this encourages a lot of depth sensors based recognition methods. With the availability of 3D joint positions extracted by a real time skeleton tracking algorithm [5], a lot of researchers use these joints to build action representations. For example, a histogram of 3D joint locations (HOJ3D) is proposed to represent human postures in [6]. Gowayyed et al. [7] propose a 2D trajectory descriptor for each skeleton joint, where the 3D joint trajectory is projected into three plane, then a histogram of oriented displacements(HOD) is used to record the angles between two consecutive motion frames in each plane.

Inspired but quite different from [7], we partition moving directions of joints into m even bins according to m vectors, and introduce a histogram of 3D directions. The histogram records the moving trend of each joint over the entire sequence. Moreover, we also propose a sequenced motion feature by extracting the geometry property of each joint. The final feature descriptor is the concatenation of these two types of features. Contributions of this paper are as follows.

1) A new histogram projection method is proposed to extract the 3D moving trend of each joint, which can describe its specific tendency in 3D space.

2) The geometry property of joints is constructed by using the relative motion of each frame with the initial status to represent the evolution of actions.

3) A novel scale-invariant skeleton joints feature descriptor based on 3D Moving Trend and Geometry (3DMTG) property, which is named as 3DMTG descriptor, is proposed for human action recognition. Experimental results show that the proposed feature descriptor has superior performance over many leading methods in the state-of-the-art, especially a better recognition ability for actions in *Cross Subject Tests*.

The remainder of this paper is organized as follows: Section II reviews related work for human action recognition. Section III introduces the process of modelling the 3DMTG feature descriptor. Section IV reports various experimental results as well as the comparison with the state-of-the-art methods. Section V summarizes the work of this paper.

II. Related Work

In recent years, there is extensive literature on depth images based human motion recognition. Depending on used feature types, these methods can be broadly divided into two categories: depth maps-based methods and skeletal joints/body parts-based methods.

Depth maps-based methods mainly extract space features along time [8]. Some authors [9][10] project depth images onto three 2D orthogonal planes to capture action features from diverse viewpoints. In [9], depth motion map (DMM) is generated by accumulating motion energy over the whole sequence and the histogram of gradient (HOG) for each DMM is computed to describe actions. Local interest points and occupancy patterns are also presented as descriptors of actions [11][12]. Vieiral *et al* [12] apply space-time occupancy patterns (STOP), where the depth map sequence is represented as a 4D grid with same-size cells whose occupancy value are recorded. A saturation scheme is used to enhance the cells containing more information about either silhouettes or moving parts of the body. In [13], the 4D spatio-temporal feature is captured using information from both RGB and depth images within a 4D cubiod, and then gradients of each cubiod along x, y, t directions are computed and concatenated as the feature representation. The dynamic bag-of-words is developed to distinguish unfinished activities by a probabilisitc model [14].

Unlike methods in the former category, in skeletal joints/body parts-based methods, human actions can be regarded as the time evolution of rigid segments connected by joints in space [15]. Feature descriptors in this category tend to use different joints information, such as joint locations, joint angles and geometric relationships between body parts to represent actions. In [16], the most informative joints are firstly captured within an instant time according to the mean or variance of joint angles. Human actions are represented by a concatenation of histograms of these joints and then recognized by comparing the Levenshtein distance. The differences of joints including posture, motion and offset information are combined to get new features named EigenJoints for action recognition [17]. Rayes *et al* [18] consider 15 3D joints as the feature descriptor at frame-level for motion modelling. In the descriptor, each skeleton joint used for gesture recognition is given different weights in different gesture classes based on the contribution of the joint to the particular gesture class[18] [19]. Sung *et al* [20] regard a human activity as a combination of a set of sub-activities over the whole sequences. They extract features including joint orientation, hand position and motion information from both RGB and depth images, and then use a two-layered Markov model to recognize activities. Some other researchers consider body parts as intermediate representation. In [21], joint angles between connected pairs of body parts are chosen as motion features instead of joint trajectories and then similarities between each angle with temporal evolution are used as representation of actions. Vemulapalli *et al* [15] use the rotation and translation to describe the relative 3D geometry between body parts, and model actions as curves in the Lie group. In this paper, we model a novel scale-invariant 3DMTG feature descriptor for human action recognition by using the skeleton joints extracted by the method proposed in [5].

III. Proposed Method

In this section, we introduce the proposed 3DMTG descriptor for novel skeleton joints presentation based on 3D Moving Trend and Geometry property. Firstly, a histogram of 26 bins is used to record the moving directions of consecutive frames over the whole trajectories in 3D space to represent the moving trend of each joint. Secondly, the geometry property of joints in each frame is modelled by the relative motion information. Finally, the 3DMTG feature descriptor is constructed by combining the two features together for action recognition.

A. *3D Moving Trend Feature*

Unlike [7] where three 2D trajectory descriptors is used to represent the 3D joints trajectory feature, we propose to utilize 3D moving directions directly. We describe 3D moving trend feature through the whole sequence to record specific tendency for each body joint. Fig. 1 shows an illustration of 3D moving trend feature modelling. The moving directions of body joints in 3D space are various while actors performing different actions, therefore, we partition 3D moving directions into m bins as shown in Fig.1 (b) (we take $m = 26$ in our experiment), and then a histogram including m bins is built to describe the moving trend feature of joints in spatial domain (as shown in Fig. 1(c)).

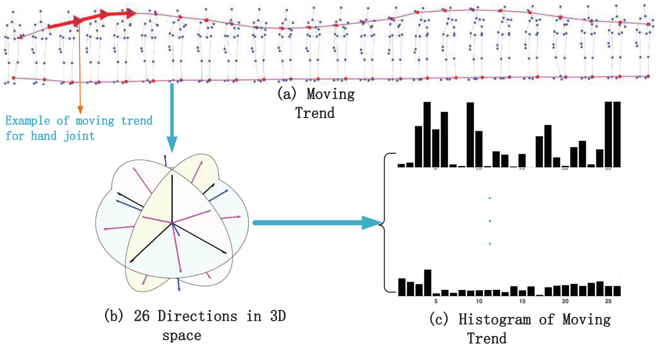

(a) Moving Trend
Example of moving trend for hand joint
(b) 26 Directions in 3D space
(c) Histogram of Moving Trend
5
10
15
20
25

Fig. 1. An illustration of 3D moving trend feature modelling. (a) 3D moving directions (red lines are moving trend of example joints and red vectors are moving directions between consecutive frames). (b) 26 directions in 3D space. The black vectors are $\mathbf{v_1}$-$\mathbf{v_6}$, the blue vectors are $\mathbf{v_7}$-$\mathbf{v_{14}}$ and the purple vectors are $\mathbf{v_{15}}$-$\mathbf{v_{26}}$ (c) Histograms of moving trend for 20 joints.

Let $V = [\mathbf{v_1}, \mathbf{v_2}, ..., \mathbf{v_m}]$ be the matrix of m vectors in 3D space. These vectors are given by:

$$\begin{array}{lll}
\mathbf{v_1} = (0,0,1)^T, & \mathbf{v_2} = (0,0,-1)^T, & \mathbf{v_3} = (0,1,0)^T, \\
\mathbf{v_4} = (0,-1,0)^T, & \mathbf{v_5} = (1,0,0)^T, & \mathbf{v_6} = (-1,0,0)^T, \\
\mathbf{v_7} = (1,1,1)^T, & \mathbf{v_8} = (-1,-1,-1)^T, & \mathbf{v_9} = (1,1,-1)^T, \\
\mathbf{v_{10}} = (-1,-1,1)^T, & \mathbf{v_{11}} = (1,-1,1)^T, & \mathbf{v_{12}} = (-1,1,-1)^T, \\
\mathbf{v_{13}} = (1,-1,-1)^T, & \mathbf{v_{14}} = (-1,1,1)^T, & \mathbf{v_{15}} = (1,1,0)^T, \\
\mathbf{v_{16}} = (-1,-1,0)^T, & \mathbf{v_{17}} = (1,-1,0)^T, & \mathbf{v_{18}} = (-1,1,0)^T, \\
\mathbf{v_{19}} = (-1,0,-1)^T, & \mathbf{v_{20}} = (1,0,1)^T, & \mathbf{v_{21}} = (1,0,-1)^T, \\
\mathbf{v_{22}} = (-1,0,1)^T, & \mathbf{v_{23}} = (0,1,1)^T, & \mathbf{v_{24}} = (0,-1,-1)^T, \\
\mathbf{v_{25}} = (0,1,-1)^T, & \mathbf{v_{26}} = (0,-1,1)^T &
\end{array} \tag{1}$$

For $i-th$ joint, given a point set:

$$P^i = \{p^i_1, ..., p^i_t, ..., p^i_F\} \tag{2}$$

where F is the length of action sequence, and t represents the time. Since p^i_t includes three coordinates x, y, z. We get the 3D direction vector $\mathbf{v^i_t}$ of $i\ th$ joint from p^i_t and p^i_{t-1}:

$$\mathbf{v^i_t} = \{x_{p^i_t} - x_{p^i_{t-1}}, y_{p^i_t} - y_{p^i_{t-1}}, z_{p^i_t} - z_{p^i_{t-1}}\} \tag{3}$$

and then calculate the $cos\langle \mathbf{v^i_t}, \mathbf{v_j} \rangle$ of angle $\theta^i(t)$ between $\mathbf{v^i_t}$ and m vectors:

$$cos\theta^i_j(t) = \frac{\mathbf{v_j} \cdot \mathbf{v^i_t}}{\|\mathbf{v^i_t}\| \|\mathbf{v_j}\|}, j \in [1, m] \tag{4}$$

where $\mathbf{v_j} \in \mathbf{V}$. Since the greater the $cos\theta^i_j(t)$ value, the more similar the direction, we use the cosine similarity $cos\theta^i_j(t)$ to describe the similarity between $\mathbf{v^i_t}$ and $\mathbf{v_j}$. In our experiment, we choose two bins that have the most similar directions (corresponding to $cos\theta^i_{first}(t)$ and $cos\theta^i_{second}(t)$ respectively) to reflect the most possible moving directions for current motion of $i-th$ joint.

$$\left\{\begin{array}{lcl}\cos\theta^{i}_{first}(t) & = & max\{cos\theta^{i}_{j}(t)\}, j \in (1,m)\\ \cos\theta^{i}_{second}(t) & = & max\{cos\theta^{i}_{j}(t)\}, j \neq first\end{array}\right. \quad (5)$$

The product of displacement and $cos\theta^{i}_{first}(t)$ and the product of displacement and $cos\theta^{i}_{second}(t)$ are finally added to the corresponding bins:

$$\left\{\begin{array}{l}bin_{first} = bin_{first} + Dis^{i}(t) \times cos\theta^{i}_{first}(t)\\ bin_{second} = bin_{second} + Dis^{i}(t) \times cos\theta^{i}_{second}(t)\end{array}\right. \quad (6)$$

where bin_{first} and bin_{second} are the corresponding bins in the histogram of 3D moving directions, $Dis^{i}(t) = \|\mathbf{v^{i}_{t}}\|$.

B. Geometry Property

This paper regards the human action as the relative movement of different body joints to the hip-center joint of human. To remove the coordinate difference caused by various distances between actors and the depth sensor, we translate the world coordinate from the depth sensor to the center of actors in each frame. This can be simply accomplished by subtracting the coordinate of hip-center point for every joint in each frame. Although the world coordinate of each frame may differ under current strategy, the advantage is obvious as hip-center point is relatively stable in majority of actions.

Apart from the feature of hip-center point relative movement, it should be noted that different actors might have different initial poses for the same action. In order to eliminate the influence of different initial poses for the rest 19 joints, this paper uses the displacement between the relative joints in current frame and the joints in the initial frame to reflect the geometry property in current frame.

Furthermore the action recognition performance will also be affected by the various body sizes of the actors. This is caused by internal difference of human or various distances between actors and the depth sensor. To solve this problem, a feature normalization method is performed on the extracted geometry property feature. Thus the proposed geometry property feature is scale-invariant to the different body sizes. The detailed description of the geometry property feature is as follows.

The movement of a point can be regarded as the composition of movements of x, y, z axes. We describe the motion property of 20 body joints separately. In each single frame, the relative motion of each joint to its initial status in three axes is recorded. Each frame represents a body pose that can be described by the locations of 20 joints.

$$I_{t} = \{p^{1}_{t}, p^{2}_{t}, ..., p^{N}_{t}\} \quad (7)$$

where N is the number of joints, p^{i}_{t} is the position of the $i-th$ joint at time t and it contains 3D coordinates x^{i}_{t}, y^{i}_{t} and z^{i}_{t}. The difference along three axes of each joint can be computed between the initial status and the current status.

In this paper, we translate the world coordinate system to the hip-center joint for each frame. The transformed coordinates of skeleton joints are as follows.

$$p^{ri}_{t} = p^{i}_{t} - p^{hipcenter}_{t}, i = 1, 2, ..., N \quad (8)$$

where p^{ri}_{t} is the relative position of the $i-th$ joint in time t. So the transformed coordinates of the frame is $I_{rt} = \{p^{r1}_{t}, p^{r2}_{t}, ..., p^{rN}_{t}\}$ and we define the geometry property of each joint in frame t as:

$$\left\{\begin{array}{l}\triangle x^{i}_{t} = x^{ri}_{t} - x^{ri}_{1},\\ \triangle y^{i}_{t} = y^{ri}_{t} - y^{ri}_{1},\\ \triangle z^{i}_{t} = z^{ri}_{t} - z^{ri}_{1},\end{array}\right. \quad (9)$$

where $(x^{ri}_{1}, y^{ri}_{1}, z^{ri}_{1}) and (x^{ri}_{t}, y^{ri}_{t}, z^{ri}_{t})$ are the three transformed coordinates of the initial status and current status, respectively. The relative displacement of the $i-th$ joint in frame t is $\triangle d^{i}_{t} : (\triangle x^{i}_{t}, \triangle y^{i}_{t}, \triangle z^{i}_{t})$, and the geometric property of current frame is:

$$g(t) = \{\triangle d^{1}_{t}, ..., \triangle d^{N}_{t}\} \quad (10)$$

We use $G(k) = \{g(1), ..., g(F)\}$ to denote the feature of action k. So the dimension of defined geometric property feature for one frame is 20×3. Although [17] also uses the difference of the joints between current frame and the initial frame, the geometric property feature defined in this paper is totally different. In [17] different combination of the joints is used and the final dimensions for each frame is 400×3, which is 20 times larger than our feature dimensions.

The length of action sequences may differ in each action instance, and this will lead to unequal length of geometry property feature. Therefore, we use the cubic spline interpolation [15] to rescale the feature before integrating them into the feature descriptor.

Finally, to acquire the scale-invariant feature for the different body sizes, we use the following normalization method.

$$G(k) = \frac{G(k)}{\parallel G(k) \parallel} \quad (11)$$

where $G(k)$ stands for the extracted geometric property feature for action k.

C. 3DMTG Feature Descriptor

This paper proposes a 3DMTG feature descriptor that is a combination of the 3D moving trend feature and geometry property feature to represent the motion information in action sequences. The general framework of the proposed 3DMTG feature descriptor is shown in Fig. 2.

The upper part of Fig. 2 is the 3D moving trend feature where a histogram of 26 bins corresponding to 3D moving directions is adopted to store the moving trend of each joint through the whole action video. The lower part of Fig. 2 is the geometry property feature which is acquired from the N frames of the action sequence. In the geometry property feature, the world coordinate is firstly translated into hip-center using Eq.(8) and the relative displacement of each joint is computed by using Eq.(9). To address unequal length of geometry property feature caused by length of action sequences, the relative displacement property of each action instance is interpolated to the unified dimension, $M \times 20 \times 3$. Both 3D moving trend and geometry property features are normalized.

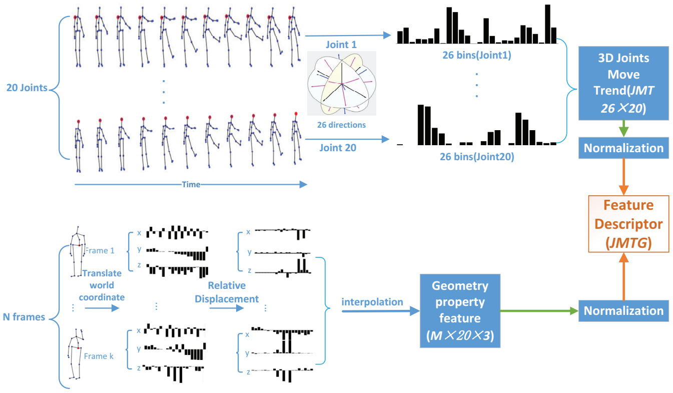


Fig. 2. An overview of the proposed 3DMTG feature descriptor.

The final 3DMTG feature descriptor for the motion is a concatenation of 3D moving trend and the geometry property. We focus on describing motion property of each single joint. The 3D moving trend feature reflects spatial motion direction of each joint in a action sequence, while the geometry property feature indicates the temporal movement of each joint. Our method builds features for joints of different body parts, so it can differentiate partial similar actions. After creating the descriptor, a linear SVM [22] classification algorithm is used for action recognition.

IV. Experiments

The proposed 3DMTG feature descriptor is evaluated for human action recognition on two publically available datasets: *MSR-Action3D* [23] dataset and *Florence3D-Action* dataset [24].

A. Dataset

1) MSR-Action3D: The *MSR-Action3D* [23] dataset has 20 action types, 10 subjects, and each subject performs each action for two or three times. The actions are *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 pickup throw*.

Similar to [23], the data is divided into three action sets *AS1*, *AS2* and *AS3*, as show in Table I. Actions with similar movement are grouped in the *AS1* and *AS2* sets, while complex actions are grouped in *AS3* set. Each set has eight actions with some overlaps between action sets. Our method adopts the 3D joint positions extracted by a real time skeleton tracking algorithm [5] to build action representation.

In each action set, there are three tests with different settings of training and testing samples: *Test One* (one third of the samples are used for training), *Test Two* (two third of the samples are used for training) and *Cross Subject Test* (samples from half of subjects are used for training).

2) Florence3D-Action: The *Florence3D-Action* dataset [24] includes nine actions performed by ten subjects for two or three times. The actions are: *wave, drink from a bottle, answer phone, clap, tight lace, sit down, stand up, read watch and bow*. Most of these actions have great similarity, for example, both *answer phone* and *drink a bottle* include a hand picking up the object to around the head.

TABLE I
Three action sets of *MSR-Action3D* dataset.

AS1	*AS2*	*AS3*
Horizontal Wave	High Wave	High Throw
Hammer	Hand Catch	Forward Kick
Forward Punch	Draw X	Side Kick
High Throw	Draw Tick	Jogging
Hand Clap	Draw Circle	Tennis Swing
Bend	Hands Wave	Tennis Serve
Tennis Serve	Forward Kick	Golf Swing
Pickup & Throw	Side Boxing	Pickup & Throw

B. Experimental settings

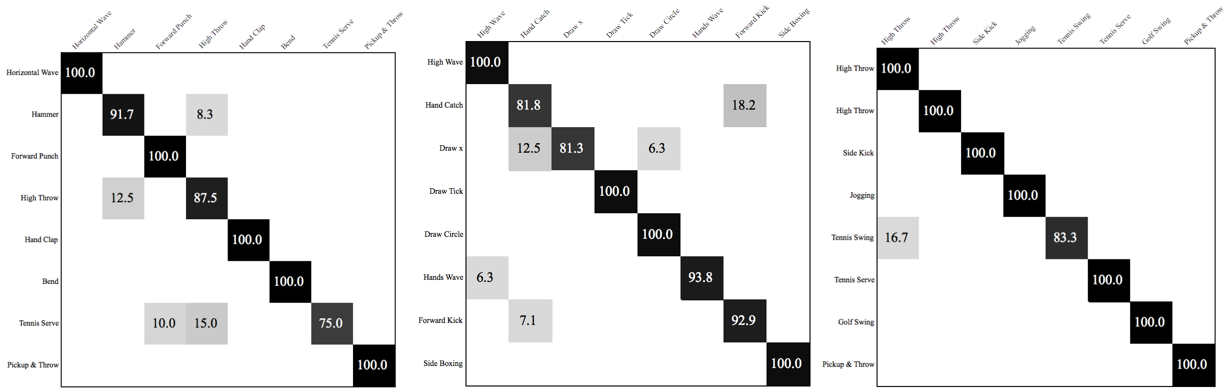
For *MSR-Action3D* dataset, although it is clear that how many samples are used in three sets, which 1/3 or 2/3 instances or which half of subjects for training is ambiguous. In order to facilitate a fair comparison with the existing methods, we consider three settings for *Test One* and *Test two* while two settings for *Cross Subject Test*. For simplicity, we define a term $eaes$ to represent that each action is performed by each subject.
Setting1: in *Test One*, the first instance of $eaes$ is for training and the rest are for testing; in *Test Two*, the first and second instances of $eaes$ are for training and the rest are for testing.
Setting2: in *Test One*, the second instance of $eaes$ is for training and the rest are for testing; in *Test Two*, the second and third instances of $eaes$ are for training and the rest are for testing.
Setting3: in *Test One*, the third instance of $eaes$ is for training and the rest are for testing; in *Test Two*, the first and third instances of $eaes$ are for training and the rest are for testing.
Setting4: in *Cross Subject Test*, samples of subjects 1, 3, 5, 7 and 9 are for training, and samples of subjects 2, 4, 6, 8 and 10 are for testing.
Setting5: in *Cross Subject Test*, samples of subjects 1, 2, 3, 4 and 5 are for training, and samples of subjects 6, 7, 8, 9 and 10 are for testing.

The selected samples of *Setting4* and *Setting5* for *Cross Subject Test* are the same as those selected in [23] and [31] respectively.

For *Florence3D-Action* dataset, we operate the *Cross Subject Test* and follow the test setting of [15], where samples from half of subjects are used for training and the rest are used for testing.

C. Result and Discussion

For *MSR-Action3D* dataset, Fig. 3 shows the confusion matrices of our 3DMTG method for *AS1*, *AS2* and *AS3* on *Cross Subject Test* . It can be seen that most actions can be 100% recognized by the proposed descriptor, especially for *AS3* that all actions except *Tennis Swing* are correctly recognized. Because actions in *AS1* and *AS2* have big intra-class variations, some actions are confused with others, such as *Hammer* and *High Throw*, *Tennis Serve* and *Forward*



(a) AS1 Cross Subject Test

(b) AS2 Cross Subject Test

(c) AS3 Cross Subject Test

Fig. 3. Confusion Matrixes of the proposed 3DMTG feature descriptor: *AS1*,*AS2* and *AS3*.

TABLE II
Recognition Accuracy (%) of *Test One* and *Test Two* on the *MSR-Action3D*.

		Test One				*Test Two*			
Method		*AS1*	*AS2*	*AS3*	**Average**	*AS1*	*AS2*	*AS3*	**Average**
state-of-the-art	Bag of 3D Points [23]	89.5	89.0	96.3	91.6	93.4	92.9	96.3	94.2
	DMM-HOG[9]	97.3	92.2	**98.0**	95.8	98.7	94.7	98.7	97.40
	STOP(skeleton)[25]	**98.2**	94.8	97.4	96.8	**99.1**	97.0	98.7	98.3
	HOJ3D[6]	98.5	96.7	93.5	96.2	98.6	97.9	94.9	97.2
	EigenJoints[17]	94.7	95.4	97.3	95.8	97.3	98.7	97.3	97.8
3DMTG	*Setting 1*	96.3	91.5	96.5	94.8	98.5	96.0	**100**	98.2
	Setting 2	**96.3**	**98.0**	**97.9**	**97.4**	**98.6**	98.7	98.6	**98.6**
	Setting 3	91.4	95.4	95.9	94.2	98.5	**98.7**	98.7	98.6
	Average	94.7	95.0	96.8	95.5	98.5	97.8	99.1	98.5

TABLE III
Average accuracy of *Cross Subject Test* on the *MSR-Action3D*.(1 - silhouette based methods, 2 - local interest points methods, 3 - skeleton joints based methods)

Setting 4 (1,3,5,7,9 subjects as training)					
	Method	*AS1*	*AS2*	*AS3*	**Average(%)**
1	Bag of 3D Points[23]	72.9	71.9	79.2	74.7
	DMM-HOG[9]	96.2	84.1	94.6	91.6
	SNV[26]	-	-	-	93.1
2	STOP [25]	91.7	72.2	98.6	87.5
	ROP [11]	-	-	-	86.5
	DSTIP [27]	-	-	-	89.3
3	HOJ3D[6]	72.9	85.5	63.5	79.0
	EigenJoints[17]	74.5	76.1	96.4	82.3
	Actionlets Ensemble [28]	-	-	-	88.2
	HOD [7]	92.4	90.2	91.4	91.3
	Vemulapalli et al.[15]	95.3	83.8	98.2	92.5
	3DMTG	**92.4**	**93.8**	**97.1**	**94.4**
Setting 5 (1,2,3,4,5 subjects as training)					
	HON4D [29]	-	-	-	88.9
	pose set[30]	-	-	-	90.2
	Moving Pose [31]	-	-	-	91.3
	3DMTG	**87.50**	**95.8**	**94.7**	**92.7**

TABLE IV
Average accuracy of *Cross Subject Test* on the *Florence3D-Action*.

Multi-Part Bag-of-Poses	82.0
Vemulapalli et al.[15]	90.9
3DMTG	**91.3**

Punch. As a result, the recognition accuracies of these actions are lower than those actions with small or none intra-class variations.

To show the good performance of the 3DMTG method, it is also compared with many leading methods in the state-of-the-arts. We compare our method with silhouette-based [9] [23] [26] [29], local interest points-based [11] [25] [27] and skeleton-based [6] [15] [17] [28] [30] [31] action recognition methods under different settings. Table II, Table III and Table IV list the average accuracies of different settings.

In Table II, the best results are highlighted in bold. The results demonstrate that our 3DMTG descriptor performs better in most cases and it even achieves 100% in *Test Two* under *setting 1*. Especially, all the recognition accuracies of our method for *Test One* and *Test Two* are better than the accuracies of skeleton-based methods (HOJ3D[6] and EigenJoints[17]).

In addition, we compare the recognition performance of our 3DMTG feature descriptor to the state-of-the-art feature descriptors on *Cross Subject Test*. The results on *MSR-Action3D* dataset and *Florence3D-Action* dataset are shown in Table III and Table IV, respectively. For *MSR-Action3D*, our method achieves accuracies over 90% for *AS1*, *AS2* and *AS3* with samples of subjects 1, 3, 5, 7 and 9 for training (*setting 4*) and the accuracies are around 95% for *AS2* and *AS3* corresponding to samples of subjects 1, 2, 3, 4 and 5 for training (*setting 5*). The proposed 3DMTG feature descriptor outperforms silhouette-based methods. Specifically, the recognition accuracy of our descriptor is approximately 20% higher than bag of 3D points [23], 3% higher than DMM-HOG [9], and 1.3% higher than SNV [26]. Moreover, our descriptor improves the average recognition rate by 5.1% compared to the best result of local interest points-based method DSTIP [27]. Its average accuracy 92.7% under *setting 5* is also 1.4% higher than the state-of-the-arts. For *Florence3D-Action*, our 3DMTG feature descriptor performs 91.3% recognition accuracy for *Cross Subject Test*, which is higher than that of [32][15].

As aforementioned, actions in both reported datasets have large intra-class variations and small inter-class variations. Higher recognition rates on three action sets of *MSR-Action3D* and *Florence3D-Action* reflect that our feature descriptor is able to tackle actions with huge intra-class variations or various actions with great similarity. It particularly can perform superior recognition on *Cross Subject Test*.

V. Conclusion

In this paper, a scale-invariant 3DMTG skeleton joints descriptor is proposed for depth based human action recognition. A effective histogram projection method is proposed to extract the Joints Moving Trend in 3D space. In addition, the relative motion of different frames with the initial status is used to present the geometry information along the whole action sequence. By combining the two types features, the proposed feature descriptor is able to represent actions better. Experimental results on datasets *MSR-Action3D* and *Florence3D-Action* show our method achieves high recognition rates on both similar actions and complex actions. Our future work will concentrate on recognition of interactions between people and people / objects.

Acknowledgment

This work was supported in part by the EU Seventh Framework Programme (No. 611391, Development of Robot-Enhanced therapy for children with AutisM spectrum disorders (DREAM)), the National Natural Science Foundation of China (No. 61403342), and China Scholarship Council.

References

[1] J. C. Niebles and L. Fei-Fei, "A hierarchical model of shape and appearance for human action classification," *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pp. 1–8, 2007.

[2] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International journal of computer vision*, vol. 79, no. 3, pp. 299–318, 2008.

[3] N. Ikizler and P. Duygulu, "Human action recognition using distribution of oriented rectangular patches," *Human Motion–Understanding, Modeling, Capture and Animation*, pp. 271–284, 2007.

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

[5] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[6] L. Xia, C.-C. Chen, and J. Aggarwal, "View invariant human action recognition using histograms of 3d joints," *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pp. 20–27, 2012.

[7] 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," *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pp. 1351–1357, 2013.

[8] M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, and J. Gall, "A survey on human motion analysis from depth data," *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*, pp. 149–187, 2013.

[9] X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," *Proceedings of the 20th ACM international conference on Multimedia*, pp. 1057–1060, 2012.

[10] C. Chen, K. Liu, and N. Kehtarnavaz, "Real-time human action recognition based on depth motion maps," *Journal of Real-Time Image Processing*, pp. 1–9, 2013.

[11] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu, "Robust 3d action recognition with random occupancy patterns," *Computer vision–ECCV 2012*, pp. 872–885, 2012.

[12] A. W. Vieira, E. R. Nascimento, G. L. Oliveira, Z. Liu, and M. F. Campos, "Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences," pp. 252–259, 2012.

[13] H. Zhang and L. E. Parker, "4-dimensional local spatio-temporal features for human activity recognition," *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pp. 2044–2049, 2011.

[14] M. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos," *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 1036–1043, 2011.

[15] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pp. 588–595, 2014.

[16] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Sequence of the most informative joints (smij): A new representation for human skeletal action recognition," *Journal of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 24–38, 2014.

[17] X. Yang and Y. Tian, "Eigenjoints-based action recognition using naive-bayes-nearest-neighbor," *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pp. 14–19, 2012.

[18] M. Reyes, G. Domínguez, and S. Escalera, "Featureweighting in dynamic timewarping for gesture recognition in depth data," *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pp. 1182–1188, 2011.

[19] S. Celebi, A. S. Aydin, T. T. Temiz, and T. Arici, "Gesture recognition using skeleton data with weighted dynamic time warping." *VISAPP (1)*, pp. 620–625, 2013.

[20] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Human activity detection from rgbd images." *plan, activity, and intent recognition*, vol. 64, 2011.

[21] E. Ohn-Bar and M. M. Trivedi, "Joint angles similarities and hog2 for action recognition," *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, pp. 465–470, 2013.

[22] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

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

[24] "http://www.micc.unifi.it/vim/datasets/3dactions/," *Florence 3D action dataset*, 2013.

[25] A. W. Vieira, E. R. Nascimento, G. L. Oliveira, Z. Liu, and M. F. Campos, "On the improvement of human action recognition from depth map sequences using space–time occupancy patterns," *Pattern Recognition Letters*, vol. 36, pp. 221–227, 2014.

[26] X. Yang and Y. Tian, "Super normal vector for activity recognition using depth sequences," *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pp. 804–811, 2014.

[27] L. Xia and J. Aggarwal, "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera," *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pp. 2834–2841, 2013.

[28] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 1290–1297, 2012.

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

[30] C. Wang, Y. Wang, and A. L. Yuille, "An approach to pose-based action recognition," *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pp. 915–922, 2013.

[31] M. Zanfir, M. Leordeanu, and C. Sminchisescu, "The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection," *Computer Vision (ICCV), 2013 IEEE International Conference on*, pp. 2752–2759, 2013.

[32] L. Seidenari, V. Varano, S. Berretti, A. Del Bimbo, and P. Pala, "Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses," *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, pp. 479–485, 2013.