Signal, Image and Video Processing (2020) 14:1227–1234
https://doi.org/10.1007/s11760-020-01644-0

ORIGINAL PAPER

Rotation-based spatial–temporal feature learning from skeleton sequences for action recognition

Xing Liu[1] · Yanshan Li[1] · Rongjie Xia[1]

Received: 3 September 2019 / Revised: 9 December 2019 / Accepted: 21 January 2020 / Published online: 2 March 2020

Abstract
Recently, skeleton-based action recognition approaches achieve great improvement by providing various features from skeleton sequences as the inputs of deep neural networks for classification. Compared with other works which directly utilize the locations of joints as the representation of each action for recognition, our paper describes the relative relations embedded in skeletons to alleviate the impact of viewpoint diversity. Specifically, we propose a novel feature descriptor based on the rotation relations in skeletons to represent a certain action. Geometric algebra (GA) is introduced to calculate and derive the rotation relations according to the particular operator, namely the rotor in GA. In order to exploit the spatial and temporal characteristics in one skeleton sequence, we design two different rotor-based feature descriptors, respectively, from one frame of skeleton and two skeletons of consecutive frames. Then an efficient feature encoding strategy is proposed to transform each kind of feature descriptor into a RGB image. Afterward, we propose a two-stream convolutional neural network(CNN) based framework to learn the RGB images generated by each skeleton sequence and then fuse the scores of two networks to provide final recognition accuracy. Extensive experimental results on NTU RGB+D, Northwestern-UCLA, Gaming 3D, SYSU and UTD-MHAD datasets have demonstrated the superiority of our method.

Keywords Action recognition · Skeleton sequence · Rotation-based feature · Spatial–temporal feature · Geometric algebra

1 Introduction

In earlier works on human action recognition, traditional approaches generally recognize actions from RGB data. Compared with RGB data, the skeleton data are not only insensitive to the background noise and illumination variations but also to the high-level representation of human action. As a result, more and more skeleton-based methods have been proposed and utilized for action recognition [1–5].

Recently, most researches mainly utilize CNN and RNN (recurrent neural network) to extract deep feature from skeleton data. The CNN-based methods are more effective to achieve promising results due to the powerful ability to capture the relationships both of the neighboring frames and long-term frames [6]. However, lots of works limit their inputs of CNN to coordinates of joints, which usually lose important structural information in the process of encoding skeleton data.

For skeleton-based action recognition, the skeleton data captured from different viewpoints make the action recognition very challenging. Hence, directly utilizing the absolute position of joints as the representation of a certain action may cause the viewpoint mismatch problem between training and testing samples, which will reduce the recognition accuracy. Consequently, our paper takes advantage of the relative geometric relations in skeleton sequence, which specifically are the rotation relations between two bones from skeleton structure. Then the rotation-based spatial–temporal feature learning method is proposed to fully exploit the spatial configuration and dynamics embedded in skeleton sequences. In our paper, the rotation relations are calculated and derived in geometric algebra (GA) space due to its simple and powerful geometric operations on subspaces and vectors [7]. Particularly, the rotor as the rotation operator in GA is utilized to calculate the rotation relations in skeletons. Then three angles are derived and defined as the rotor-based feature descriptor. Our paper proposes rotor-based spatial and temporal feature descriptors for one action, which are aimed to

✉ Yanshan Li
lys@szu.edu.cn

[1] ATR National Key Laboratory of Defense Technology, Shenzhen University, Shenzhen, China

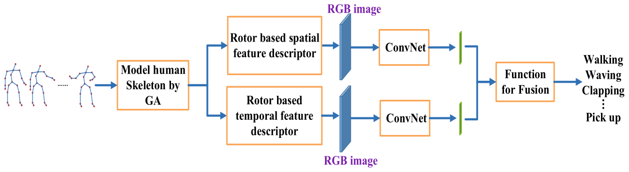

Model human Skeleton by GA
Rotor based spatial feature descriptor
Rotor based temporal feature descriptor
RGB image
RGB image
ConvNet
ConvNet
Function for Fusion
Walking
Waving
Clapping
⋮
Pick up

Fig. 1 Flowchart of our proposed method

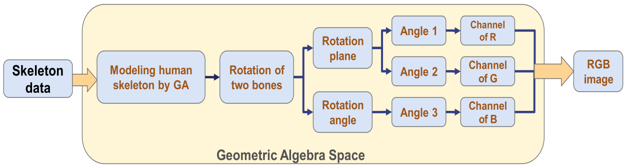


Fig. 2 The pipeline of generating a RGB image by rotor-based feature descriptor

exploit the rotation relations in one frame of skeleton and two consecutive frames of skeletons respectively. Besides, we propose an efficient feature encoding method to transform the feature descriptors into RGB images. Finally, we design two-stream CNN, and one CNN is utilized to learn rotation-based spatial feature descriptor, and the other one is used to learn rotation-based temporal feature descriptor. Thus, both the temporal and temporal characteristics of one skeletal sequence are modeled by our two-stream CNN. We fuse the classification scores from each stream to obtain the final recognition accuracy. The pipeline of our work is briefly illustrated in Fig. 1. Overall, our contributions can be summarized as follows: (1) A novel feature descriptor based on the rotation relations in skeletons by geometric algebra is proposed. (2) We propose two rotor-based feature descriptors to fully exploit the spatial and temporal patterns in skeleton sequence. (3) Simulation experiments on four challenging datasets reveal the validation of our method.

2 Related work

In this section, we briefly review the existing works that closely relate to the proposed method. In related works on skeleton-based action recognition methods, researchers have already applied the efficient geometric features derived from skeleton data for action recognition [1,2,8,9]. Ferda [10] employed the angles between two adjacent bones as the features to select most informative joints for classification. Vemulapalli [1] utilized the rotational relations of two body parts as the geometric features in Lie group. Huang [2] extended the work [1] and incorporated the Lie group structure into a deep network to learn the rotation-based features. Zhang [8,9] explored and summarized various kinds of geometric features such as position, angle and distance between two different body parts to compare the performance of those features. Specially, Li [11] proposed view-invariant shape motion representations from skeleton sequences using geometric algebra. Recently, more and more literatures have adopted CNN to learn skeleton features and achieved impressive performance [3–5]. Ke [3] transformed each skeleton sequence into three clips and used CNN to learn the spatial temporal feature. Liu [4] utilized CNN to extract robust and discriminative features from color images generated by skeleton data. Besides, Hou [12] used a standard ConvNet architecture to learn the discriminative features from skeleton sequences. Tang [5] employed the graph-based CNN to capture the dependency between the joints and achieved very competitive results compared with other state-of-the-art methods.

3 Proposed method

The proposed method consists of three parts, which is able to utilize the rotation relations in skeleton sequences to generate a RGB image for classification. The pipeline of generating a RGB image from skeleton data is given in Fig. 2.

3.1 Description of human skeleton by GA

通过GA描述人类骨架

Human actions can be denoted as movements of skeletons and each skeleton consists of hinged joints and rigid bones. The skeleton joints can be treated as points in 3D Euclidean space, and those points are able to generate various geometric entities as the representations of human skeleton. It is acknowledged that geometric algebra is well suited to solve those intuitively geometric problems. Hence, we describe the human skeleton as a geometric model in 3D geometric algebra space.

In GA, a noncommutative product called geometric product is introduced. For two vectors a and b, the geometric product ab is expressed as the sum of its inner product $a \cdot b$ and outer product $a \wedge b$:

$$ab = a \cdot b + a \wedge b \tag{1}$$

The inner product is the dot product which produces a scalar, while the outer product produces a new quantity called bivector which represents an oriented and limited plane. Since the coordinates of joints are obtained in the 3D Euclidean space $\mathbb{R}^3$, we define the 3D GA space as G^3. G^3 has an orthonormal vector basis given by $\{e_1, e_2, e_3\}$, which can represent the scalars, vector, bivector, trivector by different sets [7]

$$\{1, (e_1, e_2, e_3), (e_1 \wedge e_2, e_2 \wedge e_3, e_3 \wedge e_1), (e_1 \wedge e_2 \wedge e_3)\} \tag{2}$$

where

$$e_i^2 = 1, e_i \cdot e_j = 0, i, j = 1, 2, 3, i \neq j \tag{3}$$

Then we denote the ith skeleton joint at the fth frame for a given skeleton sequence with F frames as J_i^f. The 3D coordinates of J_i^f are $p_i^f = \left(x_i^f, y_i^f, z_i^f\right), i = 1, 2, \ldots, N, f = 1, 2, \ldots, F$. N is the total number of joints. The joint J_i^f in G^3 can be described by the basis:

$$J_i^f = x_i^f e_1 + y_i^f e_2 + z_i^f e_3 \tag{4}$$

The bone B_{ij}^f which connects two adjacent joints J_i^f and J_j^f is described as a vector:

$$B_{ij}^f = \left(x_i^f - x_j^f\right) e_1 + \left(y_i^f - y_j^f\right) e_2 + \left(z_i^f - z_j^f\right) e_3 \tag{5}$$

Based on above definitions, the descriptor to represent the relative spatial information embedded in skeleton of each frame will be derived in the following part.

3.2 Derivation of rotor-based feature descriptor

First the derivation to represent the rotation is given and then two kinds of rotor-based feature descriptors are proposed. In G^3 for any geometric entity Q, the rotation is realized by the rotor R and the rotation is then defined as

$$Q' = RQ\tilde{R} \tag{6}$$

逆转

where $\tilde{R}$ is the reverse of R and in G^3 R is given as:

$$R = \cos\left(\frac{\vartheta}{2}\right) - \sin\left(\frac{\vartheta}{2}\right) H = e^{-\frac{\vartheta}{2}H} \tag{7}$$

二维向量 旋转平面

Here H is a unit bivector in G^3. R representing the plane of the rotation and ϑ is a scalar representing the rotation angle.

In this paper, we express the bivector H according to Eq. (2)

$$H = ae_1 \wedge e_2 + be_2 \wedge e_3 + ce_3 \wedge e_1 \tag{8}$$

标量

a, b, c are all the scalars. Since H is the unit bivector which satisfies $\|H\| = a^2 + b^2 + c^2 = 1$. Due to the constraints of the scalars a, b, c, we give the following definitions aimed to get unique solution of a, b, c

$$a = \sin\psi\cos\varphi, b = \sin\psi\sin\varphi, c = \cos\psi \tag{9}$$

where $\psi \in [0, \pi], \varphi \in [0, 2\pi]$. Thus we formulate the rotation plane H by the two angles ψ and φ. Finally, the rotor R is obtained as

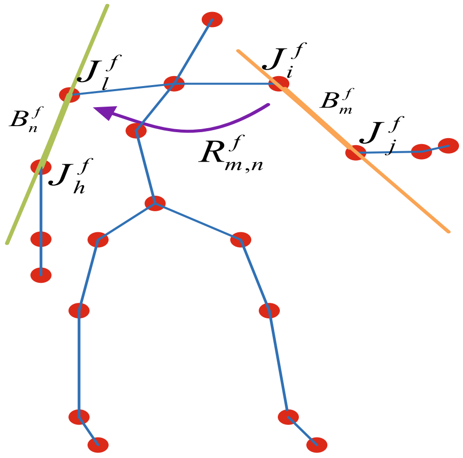


Fig. 3 The rotation from bone B_m^f to bone B_n^f

$$R = \cos\left(\frac{\vartheta}{2}\right) - \sin\left(\frac{\vartheta}{2}\right)((\sin\psi\cos\varphi\,(e_1 \wedge e_2) + \sin\psi\sin\varphi\,(e_2 \wedge e_3) + \cos\psi\,(e_3 \wedge e_1)) \tag{10}$$

推论、演绎

It can be easily deduced that the specific formulation of the rotor R is totally determined by the three angles ϑ, ψ, φ.

$$R = f(\vartheta, \psi, \varphi) \tag{11}$$

In this paper, we utilize the three angles as the rotor-based feature descriptor to represent the geometric relation of rotation.

表示旋转的几何关系。

3.2.1 Rotor-based spatial feature descriptor

In this work, we utilize the rotational relations of two bones at the same frame to represent the relative spatial information of the skeletal structure. The rotation is calculated by the rotor in GA space. Define two different bones B_m^f and B_n^f at fth frame. B_m^f is connected by two adjacent joints J_i^f and J_j^f. B_n^f consists of two adjacent joints J_l^f and J_h^f. As shown in Fig. 3, we utilize the rotational relations of two different bones at the same frame to obtain the rotor-based spatial feature descriptor.

The formulations of B_m^f and B_n^f in G^3 are described according to Eq. (5) by the 3D coordinates of corresponding adjacent joints. By carefully observing the derivation of rotor above, we can note that the norm of geometric entity for rotation is not necessary to derive the rotor. Hence, without loss of generality, we assume the norms of B_m^f and B_n^f equal to 1 to reduce the calculation complexity.

We utilize the rotor $R_{m,n}^f$ to represent the rotational relation from bone B_m^f to B_n^f in G^3. As derived above, $R_{m,n}^f$

推导转子，并不需要用于旋转的几何实体的范式

without loss of generality
不失一般性

consists of the rotation angle $\vartheta_{m,n}^{f}$ and the rotation plane $H_{m,n}^{f}$ referred to Eq. (7).

First, the rotation angle $\vartheta_{m,n}^{f}$ can be derived by the inner product of B_m^f and B_n^f:

$$B_m^f \cdot B_n^f = \left|B_m^f\right| \left|B_n^f\right| \cos \vartheta_{m,n}^{f} \tag{12}$$

Besides, the two vectors B_m^f and B_n^f are all in the rotation plane $H_{m,n}^{f}$ so that $H_{m,n}^{f}$ can be determined by the outer product of B_m^f and B_n^f. Consequently, the two vectors B_m^f and B_n^f construct a bivector representing the rotation plane $H_{m,n}^{f}$:

$$H_{m,n}^{f} = B_m^f \wedge B_n^f \tag{13}$$

According to the 3D coordinates of involved joints, we can further calculate the formulation of $H_{m,n}^{f}$:

$$\begin{aligned} H_{m,n}^{f} = B_m^f \wedge B_n^f = \frac{1}{N_{ij}^{f} \cdot N_{lh}^{f}} \\ \Big\{\Big[\left(x_i^f - x_j^f\right)\left(y_l^f - y_h^f\right) - \left(y_i^f - y_j^f\right)\left(x_l^f - x_h^f\right)\Big] e_1 \wedge e_2 \\ + \Big[\left(y_i^f - y_j^f\right)\left(z_l^f - z_h^f\right) - \left(z_i^f - z_j^f\right)\left(y_l^f - y_h^f\right)\Big] e_2 \wedge e_3 \\ + \Big[\left(z_i^f - z_j^f\right)\left(x_l^f - x_h^f\right) - \left(x_i^f - x_j^f\right)\left(z_l^f - z_h^f\right)\Big] e_3 \wedge e_1\Big\} \end{aligned} \tag{14}$$

where

$$\begin{aligned} N_{ij}^{f} &= \sqrt{\left(x_i^f - x_j^f\right)^2 + \left(y_i^f - y_j^f\right)^2 + \left(z_i^f - z_j^f\right)^2}, \\ N_{lh}^{f} &= \sqrt{\left(x_l^f - x_h^f\right)^2 + \left(y_l^f - y_h^f\right)^2 + \left(z_l^f - z_h^f\right)^2} \end{aligned} \tag{15}$$

Compared with Eqs. (8), (9) and (14), we can directly compute the values of the two angles $\psi_{m,n}^{f}$, $\varphi_{m,n}^{f}$ which determine the rotation plane $H_{m,n}^{f}$.

Finally we use the three angles $\vartheta_{m,n}^{f}$, $\psi_{m,n}^{f}$, $\varphi_{m,n}^{f}$ to consist the rotor-based spatial feature descriptor $S_{m,n}^{f}$, which represents the rotation from bone B_m^f to B_n^f:

$$S_{m,n}^{f} = \left[\vartheta_{m,n}^{f}, \psi_{m,n}^{f}, \varphi_{m,n}^{f}\right]. \tag{16}$$

3.2.2 Rotor-based temporal feature descriptor

In order to describe the geometric relative relations not only in one skeleton at single frame, we propose another descriptor utilizing the rotational relations in skeletons from two consecutive frames, which is denoted as rotor-based temporal feature descriptor.

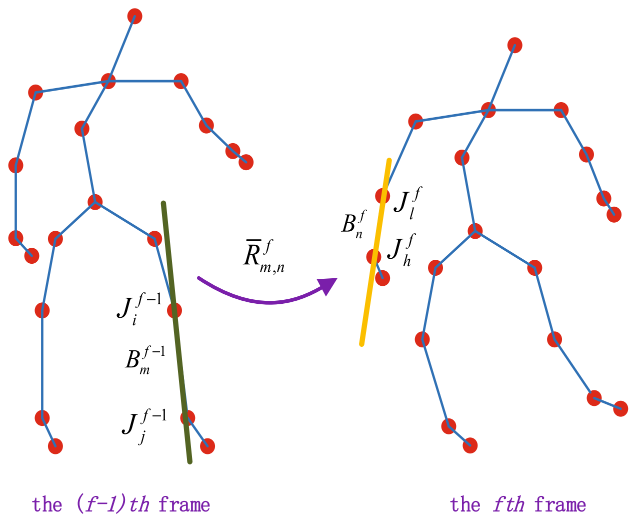


Fig. 4 Rotation of B_m^{f-1} to B_n^f from two consecutive frames

Specifically, the second proposed feature descriptor is derived in G^3 by computing the rotor between two bones, respectively, at two consecutive frames. Denote B_m^{f-1} as the bone at $(f-1)$th frame connected by two joints J_i^{f-1}, J_j^{f-1}. Meanwhile B_n^f represents the bone at fth frame connected by two joints J_l^f, J_h^f. It is noted that B_m^{f-1} and B_n^f are viewed as vectors in G^3. As depicted in Fig. 4, the rotor $\bar{R}_{m,n}^{f}$ to describe the rotational relation between two vectors B_m^{f-1}, B_n^f is calculated to derive proposed temporal feature descriptor $T_{m,n}^{f}$.

The derivation to obtain the temporal feature descriptor is similar in Sect. 3.2.1. Three angles are denoted as $\bar{\vartheta}_{m,n}^{f}$, $\bar{\psi}_{m,n}^{f}$, $\bar{\varphi}_{m,n}^{f}$ to consist $T_{m,n}^{f}$. Referring to Eqs. (8), (9), (14) and (15), we get the values of the three angles by the 3D coordinates of the involved joints.

$$T_{m,n}^{f} = \left[\bar{\vartheta}_{m,n}^{f}, \bar{\psi}_{m,n}^{f}, \bar{\varphi}_{m,n}^{f}\right]. \tag{17}$$

3.3 Encoding method to generate a RGB image

In this part, we aggregate proposed feature descriptor from all frames and all pairs of bones in one skeleton sequence together through encoding them in a RGB figure to fully represent a specific action.

Figure 5 shows the structure of generated RGB image by our proposed encoding method. The column corresponds to the fth frame of the skeleton sequence, and the row corresponds to the ith pair of two bones where the two bones in the ith pair are numbered as m and n respectively. Besides, each of the three elements $\vartheta_{m,n}^{f}$, $\psi_{m,n}^{f}$, $\varphi_{m,n}^{f}$ is represented as one of the three corresponding components (R, G, B) of a pixel in a color image. Thus, the color values of (R, G, B)

这想法真牛！

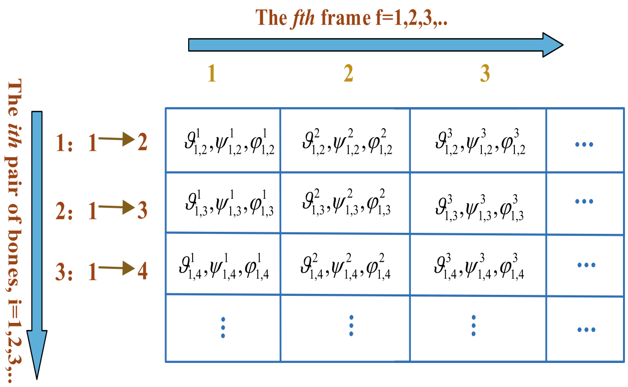

The *fth* frame f=1,2,3,..
1
2
3
The *ith* pair of bones, i=1,2,3,..
1: 1→2
2: 1→3
3: 1→4
$\vartheta^1_{1,2},\psi^1_{1,2},\varphi^1_{1,2}$
$\vartheta^2_{1,2},\psi^2_{1,2},\varphi^2_{1,2}$
$\vartheta^3_{1,2},\psi^3_{1,2},\varphi^3_{1,2}$
$\vartheta^1_{1,3},\psi^1_{1,3},\varphi^1_{1,3}$
$\vartheta^2_{1,3},\psi^2_{1,3},\varphi^2_{1,3}$
$\vartheta^3_{1,3},\psi^3_{1,3},\varphi^3_{1,3}$
$\vartheta^1_{1,4},\psi^1_{1,4},\varphi^1_{1,4}$
$\vartheta^2_{1,4},\psi^2_{1,4},\varphi^2_{1,4}$
$\vartheta^3_{1,4},\psi^3_{1,4},\varphi^3_{1,4}$
...

Fig. 5 Structure of generated RGB image

on location (i, f) are

$$(R, G, B)_{i,f} = (\vartheta^f_{m,n}, \psi^f_{m,n}, \varphi^f_{m,n}) \tag{18}$$

Considering the differences of values in 3D skeleton and image, we normalize the pixel values to be within 0-255 and each image is normalized to 224 × 224 pixels for convenience as the inputs of deep networks. As described before, two kinds of feature descriptors are proposed based on one skeleton sequence. Thus, for one action two different color images are generated and fed to deep networks for classification. Among various deep networks, CNN is used most widely in skeleton-based action recognition approaches by utilizing its powerful representation ability. In this work, we design a two-stream CNN-based network to learn deep features from two color images.

Additionally, to explore the complementary property of deep features generated from each CNN, we introduce a weighted method to fuse the results from two streams. Set two images I_s and I_t as the inputs of two streams, respectively. Then denote $prob\,(l\,|I_s)$ as the posterior probability obtained by the output of the stream with image I_s, which means the probability of image I_s belonging to the lth action class. Similarly for image I_t, the posterior probability is denoted as $prob\,(l\,|I_t)$. Thus, for a skeleton sequence S its class score is derived by proposed weighted fusing method:

$$score\,(l\,|S) = \eta_s \cdot prob\,(l\,|I_s) + \eta_t \cdot prob\,(l\,|I_t) \tag{19}$$

η_s and η_t are relative weights of the two-stream predictions and their sum is equal to 1. Finally, the label of sequence S is obtained by its class score.

4 Experiments

We have evaluated and empirically analyzed the proposed method on five common benchmark datasets for which the coordinates of skeletal joints are provided. The five datasets are NTU RGB+D, Northwestern-UCLA, Gaming 3D, SYSU-3D and UTD-MHAD. Details of the experiments and results are given as follows.

Table 1 Comparison on the NTU dataset

Method	CV (%)	CS (%)
LI-CNN (2014) [1]	52.76	50.08
Dynamic Skeletons (2017) [15]	65.22	60.23
Synthesized CNN (2017) [4]	87.21	80.03
JDM (2017) [16]	82.30	76.20
JTM (2018) [17]	81.08	76.32
ST-LSTM (Tree) + Trust Gate(2018) [18]	77.70	69.20
GCA-LSTM (2018) [19]	84.00	76.10
Clips + CNN + MTLN(2018) [20]	84.83	79.57
FO-GASTM (2019) [11]	90.05	82.83
MT-3D (2019) [21]	90.12	82.96
RSF	84.78	81.55
RTF	82.47	81.26
RSF + RTF	87.32	83.01

4.1 Implementation details

Two-stream CNNs are involved to obtain the final predication. Meanwhile the spatial and temporal feature-based images are the inputs of each network, respectively. We use ResNet-50 [13] with pretrained parameters from ImageNet for classification. The batch size is set to 32 for NTU RGB+D and 16 for other four datasets. The implementation is based on Pytorch frame with two NVIDIA GeForce GTX 1080 GPU. First we set both the weights η_s and η_t as 0.5 in our experiment. The network weights were tuned using the mini-batch stochastic gradient descent with the momentum value set to 0.9 and weight decay set to 0.0005. Learning rate is set to 0.001, and the maximum training cycle is set to 100, and the learning rate decreased every 20 cycles.

4.2 NTU RGB+D dataset

The NTU RGB+D dataset(NTU) [14] is currently the one of the largest Kinect captured datasets with 56880 video samples of skeleton data for human action recognition. It contains 60 different action classes and 40 subjects where each subject has 25 joints. Each action is captured by 3 cameras at the same height but from different horizontal angles. Two benchmarks are recommended for this dataset: (1) Cross-subject (CS): 40320 samples from 20 subjects were used for training, and the other samples for testing. (2) Cross-view (CV): samples captured from cameras 2 and 3 were used for training, while samples from camera 1 were employed for testing.

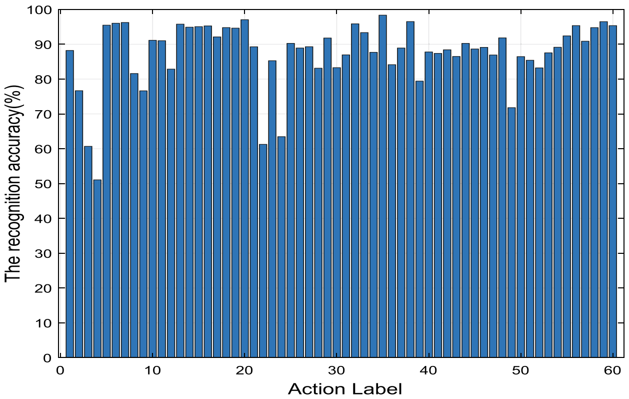


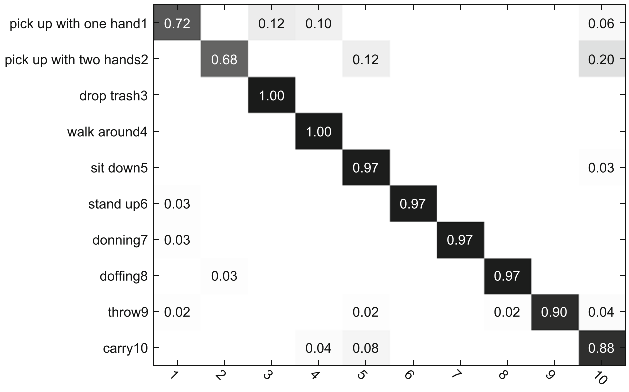
Fig. 6 The recognition accuracy (%) of each class of action on the NTU dataset for the CV setting

The results and comparison on the NTU are shown in Table 1. The accuracies of proposed rotor-based spatial feature (RSF) and rotor-based temporal feature (RTF) achieve 84.78% and 82.47%, respectively, in CV protocol. Besides, our proposed method can reach 87.32% and 83.01% after feature fusion, which outperforms most of state-of-the-art methods listed in Table 1. The results demonstrate the efficiency of our proposed approach which describes the relative geometric information from skeletons in GA. It is noted that our method achieves significant improvement compared with other works based on geometric features [2,15–17,19]. Moreover, our method outperforms some end-to-end trainable deep networks-based works [18,20]. Besides, we can observe that the works [11,21] achieve better performance than our method. The main reasons are as follows. [21] considered various geometric features of joints including the locations and relative relations from different views, while our method only utilizes one kind of geometric feature which are the rotations between two bones to represent skeletons. [11] utilized a four-stream CNN-based framework to learn different predefined active representations and a data pre-processing is implemented to realize view invariance. However, in our work only two-stream CNN is used for fusion work to obtain the final result and we explore the relative relation embedded in the human skeleton to deal with the view diversity instead of using per-processing skeletal data.

Figure 6 shows the histogram of the recognition accuracy of each action classes on NTU for the CV setting. The index of the horizontal axis denotes the Id of action as provided in [14]. It is noted that for most classes, our scheme performs very well and achieves the recognition accuracy around 90%. However, several actions such as ‘4’ denoting ‘brushing hair’ are not recognized very well because those hand actions are very similar to others.

Table 2 Comparison on the N-UCLA dataset

Method	Accuracy (%)
Ensemble TS-LSTM(2017) [24]	89.22
Multi-task RNN(2018) [23]	87.30
CVR-CNN(2018) [25]	89.38
FO-GASTM(2019) [11]	91.30
RSF	90.65
RTF	89.08
RSF+RTF	91.31



	1	2	3	4	5	6	7	8	9	10
pick up with one hand1	0.72		0.12	0.10						0.06
pick up with two hands2		0.68			0.12					0.20
drop trash3			1.00							
walk around4				1.00						
sit down5					0.97					0.03
stand up6	0.03					0.97				
donning7	0.03						0.97			
doffing8		0.03						0.97		
throw9	0.02				0.02			0.02	0.90	0.04
carry10				0.04	0.08					0.88

Fig. 7 Confusion matrix on the N-UCLA dataset

4.3 Northwestern-UCLA dataset

Northwestern-UCLA(N-UCLA) dataset [22] contains 1494 skeleton sequences of 10 action classes. It includes 10 actions performed by 10 different subjects. Each action is performed under three different views. Each skeleton subject contains 20 joints. Following the evaluation protocol in [22], we use the samples from the first two cameras as the training data and the samples from the third camera as the testing data.

The comparisons on the N-UCLA dataset are shown in Table 2. Our method achieves competitive result, and the recognition accuracy reaches 91.31%, which outperforms the previous state-of-the-art methods. The performance of our method is much superior than the results based on RNN or LSTM [23,24]. Besides, the result of our method is slightly higher than that of the four-stream CNN-based work [11], which demonstrates our proposed rotation-based descriptors are invariant to the diversity of different viewpoints.

The confusion matrix of our method on the N-UCLA dataset is shown in Fig. 7. By closely examining the matrix, we can observe that most of the activities are distinguished well, which demonstrates the efficiency of our proposed method. It can be also observed that the most significant classification error occurs between the two actions ‘pick up with two hand’ and ‘carry.’ The main reason we believe is that those actions include interactions with objects, which are

这些操作包括与对象的交互

Table 3 Comparison on the G3D dataset

Method	Accuracy (%)
ELC-KSVD(2014) [27]	82.37
LRBM(2015) [28]	90.50
LI-CNN(2017) [2]	89.10
MSD-CNN(2018) [29]	93.90
RSF	90.72
RTF	87.73
RSF+RTF	94.01

Table 4 Comparison on the SYSU dataset

Method	Accuracy (%)
VA-LSTM(2017) [30]	77.50
ST-LSTM(Tree)+Trust Gate (2018) [18]	76.50
DPRL+GCNN(2018) [5]	76.90
GCA-LSTM(2018) [19]	78.60
GGCN(2018) [31]	77.90
SR-TSL(2018) [32]	81.90
RSF	79.14
RTF	76.76
RSF+RTF	81.94

Table 5 Comparison on the UTD-MHAD dataset

Method	Accuracy (%)
Cov3DJ(2013) [34]	85.58
SOS(2018) [12]	86.97
JTM(2018) [17]	87.90
MSD-CNN(2018) [29]	96.20
HS-CNN(2019) [35]	92.40
MT-3D(2019) [21]	95.58
RSF	91.63
RTF	93.95
RSF+RTF	97.67

difficult to distinguish simply by the skeletal data of human body.

4.4 Gaming 3D dataset

subject？？？？
与actions的区别？

The Gaming 3D(G3D) dataset [26] contains 10 subjects performing 20 gaming actions and includes totally 663 skeleton sequences. We use the skeleton data from the G3D dataset which is split by subjects where the first 5 subjects are used for training and the remaining 5 subjects for testing [26].

The results and comparison on the G3D dataset are shown in Table 3. It is noted that our method achieves the recognition accuracy of 94.01% after feature fusion, which outperforms the other methods. The results demonstrate that our method performs well also on small dataset.

4.5 SYSU-3D dataset

The SYSU-3D(SYSU) dataset [15] focusing on human–object interactions includes 40 participants performing 12 different activities and contains 480 video clips. We employ the videos performed by 20 subjects for training, and the rest 20 subjects for testing. We adopt 30-fold cross-validation [15]. The mean accuracy is shown on this dataset.

The result and comparison on SYSU dataset are shown in Table 4. The mean accuracies based on RSF, RTF as well as the fusion of RSF and RTF are, respectively, given. It is noted that our proposed method outperforms the state-of-the-art approaches. As explained previously, a novel representation proposed in our work extracts the relative geometry embedded in skeletons while the absolute locations of joints are not important. Furthermore, the relatively geometric relations can directly represent the specific pose of human since the poses are different when human interact with different objects, which probably is the reason why our proposed representation outperforms others in [5,18,19,30].

4.6 UTD-MHAD dataset

UTD-MHAD [33] is a multimodal action dataset, captured by one Microsoft Kinect camera and one wearable inertial sensor. This dataset contains 27 actions performed by 8 subjects (4 females and 4 males) with each subject performing each action 4 times. After removing three corrupted sequences, the dataset has 861 sequences. For this dataset, the data from the subject numbers 1, 3, 5, 7 were used for training, and the data for the subject numbers 2, 4, 6, 8 were used for testing.

The performance comparison between our method and other state-of-the-art methods is given in Table 5. Apparently, our method outperforms other methods on the UTD-MHAD dataset.

5 Conclusions

In this paper, we investigate the efficiency of spatial and temporal feature learning based on the rotations from the skeleton sequences. We describe the human skeleton and derive the rotation between bones in GA space to generate the proposed feature descriptors. According to the process to derive the rotor-based feature descriptor, our method effectively takes the advantages of simple representation and calculation by GA. Besides, proposed feature encoding strategy is capable to easily transform the feature descriptors into RGB images fed to CNN for classification. Furthermore, the two-stream scheme in the method is very helpful to enhance the performance. Experimental results on four challenging benchmark datasets demonstrate the validation of our method. In the future, we would like to extend our framework to deal with multi-person involved activities.

Acknowledgements The authors would like to thank all reviewers' constructive advice for improving this work. This work was partially supported by National Natural Science Foundation of China (Nos. 61771319 and 61871154), Natural Science Foundation of Guangdong Province (Nos. 2017A030313343 and 2019A1515011307), Shenzhen Science and Technology Project (No. JCYJ20180507182259896).

References

1. Vemulapalli, R., Arrate, F., Chellappa, R.: Human action recognition by representing 3d skeletons as points in a lie group. In: CVPR, pp. 588–595 (2014)
2. Huang, Z., Wan, C., Probst, T.: Deep learning on lie groups for skeleton-based action recognition. In: CVPR, pp. 6099–6108 (2017)
3. Ke, Q., Bennamoun, M., An, S.: Learning clip representations for skeleton-based 3d action recognition. IEEE Trans. Image Process. **27**(4), 2842–2855 (2018)
4. Liu, M., Liu, H., Chen, C.: Enhanced skeleton visualization for view invariant human action recognition. Pattern Recognit. **68**, 346–362 (2017)
5. Tang, Y., Tian, Y., Lu, J.: Deep progressive reinforcement learning for skeleton-based action recognition. In: CVPR, pp. 5323–5332 (2018)
6. Jonas G., Michael A., David, G.: Convolutional sequence to sequence learning. In: ICML, pp. 1243–1252 (2017)
7. Dorst, L., Mann, S.: Geometric algebra: a computational framework for geometrical applications (part 1). Comput. Gr. Appl. **22**(4), 58–67 (2002)
8. Zhang, S., Liu, X., Xiao, J.: On geometric features for skeleton-based action recognition using multilayer LSTM networks. In: WACV, pp. 148–157 (2017)
9. Zhang, S., Yang, Y., Xiao, J.: Fusing geometric features for skeleton-based action recognition using multilayer LSTM networks. IEEE Trans. Multimed. **20**(9), 2330–2343 (2018)
10. Ofli, F., Chaudhry, R., Kurillo, G.: Sequence of the most informative joints (smij): a new representation for human skeletal action recognition. J. Vis. Commun. Image Represent. **25**(1), 24–38 (2014)
11. Li, Y., Xia, R., Liu, X.: Learning shape-motion representations from geometric algebra spatio-temporal model for skeleton-based action recognition. In: ICME, pp. 1066–1071 (2019)
12. Hou, Y., Li, Z., Wang, P.: Skeleton optical spectra-based action recognition using convolutional neural networks. IEEE Trans. Circuits Syst. Video Technol. **28**(3), 807–811 (2018)
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017)
14. Shahroudy A., Liu J., Ng T.-T.: Ntu rgb+d: a large scale dataset for 3d human activity analysis. In: CVPR, pp. 1010–1019 (2016)
15. Hu J.F., Zheng, W., Lai, J.: Jointly learning heterogeneous features for RGB-D activity recognition. In: CVPR, pp. 5344–5352 (2015)
16. Li, C., Hou, Y., Wang, P.: Joint distance maps based action recognition with convolutional neural networks. IEEE Signal Process. Lett. **24**(5), 624–628 (2017)
17. Wang, P., Li, Z., Hou, Y.: Action recognition based on joint trajectory maps using convolutional neural networks. Knowl. Based Syst. **158**(15), 43–53 (2018)
18. Liu, J., Shahroudy, A., Xu, D.: Skeleton-based action recognition using spatio-temporal LSTM network with trust gates. IEEE Trans. Pattern Anal. Mach. Intell. **40**(12), 3007–3021 (2018)
19. Jun, L., Gang, W., Duan, L.: Skeleton-based human action recognition with global context-aware attention LSTM networks. IEEE Trans. Image Process. **27**(4), 1586–1599 (2018)
20. Li, C., Cui, Z., Zheng, W.: Spatio-temporal graph convolution for skeleton based action recognition. In: AIAA, pp. 3482–3489 (2018)
21. Li, C., Hou, Y., Wang, P.: Multiview-based 3-d action recognition using deep networks. IEEE Trans. Hum. Mach. Syst. **49**(1), 95–104 (2019)
22. Wang, J., Nie, X., Xia, Y.: Cross-view action modeling, learning and recognition. In: CVPR, pp. 2649–2656 (2014)
23. Wang, H., Wang, L.: Learning content and style: joint action recognition and person identification from human skeletons. Pattern Recognit. **81**, 23–35 (2018)
24. Lee, I., Kim, D., Kang, S.: Ensemble deep learning for skeleton-based action recognition using temporal sliding LSTM networks. In: ICCV, pp. 1012–1020 (2017)
25. Ren, J., Reyes, N., Barczak, A.: Toward three dimensional human action recognition using a convolutional neural network with correctness-vigilant regularizer. J. Electron. Imaging **27**(4), 043040 (2018)
26. Victoria, B., Dimitrios, M., Vasileios, A.: G3d: a gaming action dataset and real time action recognition evaluation framework. In: CVPR Workshops, pp. 7–12 (2012)
27. Zhou, L., Li, W., Zhang, Y.: Discriminative key pose extraction using extended LC-KSVD for action recognition. In: DICTA, pp. 1–8 (2014)
28. Nie, S., Wang, Z., Ji, Q.: A generative restricted boltzmann machine based method for high-dimensional motion data modeling. Comput. Vis. Image Underst. **136**, 14–22 (2015)
29. Li, B., He, M., Dai, Y.: 3d skeleton based action recognition by video-domain translation-scale invariant mapping and multi-scale dilated cnn. Multimed. Tools Appl. **1**, 1–21 (2018)
30. Zhang, P., Lan, C., Xing, J.: View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In: ICCV, pp. 2136–2145 (2017)
31. Xiang G., Wei, H., Jiaxiang, T.: Generalized graph convolutional networks for skeleton-based action recognition. arXiv preprint arXiv:1811.12013 (2018)
32. Chen, S., Ya, J., Wei, W.: Skeleton-based action recognition with spatial reasoning and temporal stack learning. In: ECCV, pp. 106–121 (2018)
33. Chen, C., Jafari, R., Kehtarnavaz, N.: UTD-MHAD: a multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In: Proceedings of the IEEE International Conference on Image Processing, pp. 168–172 (2015)
34. Hussein, M.E., Torki, M., Gowayyed, M.A.: Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In IJCAI, pp. 2466–2472 (2013)
35. Ran, C., Gang, H., Aichun, Z.: Hard sample mining and learning for skeleton-based human action recognition and identification. IEEE Access **7**, 8245–8257 (2019)