




# A Novel Multi-feature Skeleton Representation for 3D Action Recognition

Lian Chen<sup>1</sup>, Ke Lu<sup>1,2</sup>, Pengcheng Gao<sup>1</sup>, Jian Xue<sup>1</sup> , and Jinbao Wang<sup>3</sup>

<sup>1</sup> University of Chinese Academy of Sciences, Beijing, China  
{chenlian17, gaopengcheng15}@mails.ucas.ac.cn  
{luk, xuejian}@ucas.ac.cn

<sup>2</sup> Peng Cheng Laboratory, Vanke Cloud City Phase I Building 8, Xili Street,  
Nanshan District, Shenzhen, China

<sup>3</sup> Southern University of Science and Technology, Shenzhen, China  
linkingring@163.com

**Abstract.** Deep-learning-based methods have been used for 3D action recognition in recent years. Methods based on recurrent neural networks (RNNs) have the advantage of modeling long-term context, but they focus mainly on temporal information and ignore the spatial relationships in each skeleton frame. In addition, it is difficult to handle a very long skeleton sequence using an RNN. Compared with an RNN, a convolutional neural network (CNN) is better able to extract spatial information. To model the temporal information of skeleton sequences and incorporate the spatial relationship in each frame efficiently using a CNN, this paper proposes a multi-feature skeleton representation for encoding features from original skeleton sequences. The relative distances between joints in each skeleton frame are computed from the original skeleton sequence, and several relative angles between the skeleton structures are computed. This useful information from the original skeleton sequence is encoded as pixels in grayscale images. To preserve more spatial relationships between input skeleton joints in these images, the skeleton joints are divided into five groups: one for the trunk and one for each arm and each leg. Relationships between joints in the same group are more relevant than those between joints in different groups. By rearranging pixels in encoded images, the joints that are mutually related in the spatial structure are adjacent in the images. The skeleton representations, composed of several grayscale images, are input to CNNs for action recognition. Experimental results demonstrate the effectiveness of the proposed method on three public 3D skeleton-based action datasets.

**Keywords:** 3D action recognition · Convolutional neural network · Deep learning · Skeleton representation

---

This work is supported by the National Key R&D Program of China (2017YFB1002-203), National Natural Science Foundation of China (62032022, 61671426, 61972375, 61871258, 61929104), Beijing Municipal Natural Science Foundation (4182071), the Fundamental Research Funds for the Central Universities (Y95401YXX2) and Scientific Research Program of Beijing Municipal Education Commission (KZ201911417048).

© Springer Nature Switzerland AG 2021

A. Del Bimbo et al. (Eds.): ICPR 2020 Workshops, LNCS 12665, pp. 365–379, 2021.

[https://doi.org/10.1007/978-3-030-68821-9\\_33](https://doi.org/10.1007/978-3-030-68821-9_33)

# 1 Introduction

In recent years, human action recognition has been applied to many fields, such as video surveillance, medical monitoring, security, intelligent houses, and content-based video retrieval [8, 10]. With the development of deep learning, action recognition has proved to be highly effective in these application areas. The modalities of input data usually include RGB videos, depth maps, and three-dimensional (3D) skeleton data. RGB is widely used in action recognition, but it can easily be affected by illumination changes and appearance of texture, resulting in ambiguity. Since the appearance of 3D cameras, such as the Microsoft Kinect, the use of skeleton data has become increasingly popular. Skeleton data are robust to variations in illumination, camera viewpoint changes, and texture variation in comparison with RGB and depth data. In this paper, we focus on 3D skeleton-based action recognition.

A skeleton data sequence consists of a series of 3D coordinates of body joints. In each skeleton sequence, the temporal information of the whole sequence describes the dynamics of action, and the spatial information in each skeleton frame describes the relationship between joints. It is of great importance to extract features from the original skeleton sequence and retain temporal and spatial information, as much as possible, for further classification. The recurrent neural network (RNN) and its extended version with long short-term memory (LSTM) [7] have been used for skeleton-based action recognition and have been shown to be effective [6, 20, 22, 24, 25, 28, 30]. They are mainly used to model the long-term context information along the temporal dimension by representing motion-based dynamics [19]. However, the RNN and LSTM lack the ability to model the entire information of a very long sequence. In addition, they focus mainly on the temporal relationship between skeleton frames, paying less attention to spatial structure in a single frame. However, the structure of joints in each skeleton frame is very important for discrimination between action categories. Not only in the field of action recognition, but also in other fields such as person re-identification, the importance of capturing and understanding information about various parts of the human body is raised. In [12], human body semantic analysis is used to extract local visual clues. In [36], the structure of the human body is used to enhance the recognition capability. [29] also proposed the importance of accurately locating each part of pedestrian and consider the continuity of information transition between each part. The convolutional neural network (CNN) [15] performs well in representation learning and has been proved effective in skeleton-based action recognition [1, 2, 4, 5, 13, 34]. Compared with an LSTM, a CNN has the advantage of learning spatial structure information. In the CNN-based method for action recognition, the manner in which the spatial relationship is extracted and the method of modeling the long-term sequence are of crucial importance.

In this paper, we propose a CNN-based method for extracting useful features from an original skeleton sequence and encoding them into grayscale images. The association between skeleton frames reflects the temporal information of the action, whereas the relationship between skeleton joints in each frame reflects the

spatial information of the action. Qihong Ke et al. [13] proposed an effective encoding method that models the temporal dynamics of the original skeleton sequence as a single image. By selecting several key joints in a single frame and calculating the relative distance between other joints and the key joints, each image reveals the connection information of the skeleton structure, and all images incorporate the spatial information of the whole sequence. In addition, the arrangement of data in the encoded image reflects the spatial connection between the joints in each skeleton frame. In this paper, we propose a new method that divides the skeleton structure into five parts and rearranges the skeleton joints. In this manner, the encoded images can better represent the relationships in the original skeleton structure. In addition, we define several important angles in a single skeleton frame. These data are encoded into an image, to incorporate more useful information and better describe the spatial features. In this manner, each original skeleton sequence can be encoded into six grayscale images: five images corresponding to five key joints and one image corresponding to the angles. These images are then input to six identical CNNs, which are trained simultaneously. Finally, we conduct score fusion to obtain the final action classification.

The main contributions of our work are as follows:

- A multi-feature skeleton representation, based on spatio-temporal information processing, is proposed.
- A useful method for space division and encoding, which corresponds to the human skeleton structure, is proposed. This can efficiently incorporate information about the relationships between adjacent joints in human movement.
- Several important joint angles, for the human skeleton structure, are defined and calculated to reflect the changes of movement and incorporate spatial motion information.

Experiments were conducted on three 3D skeleton-based action recognition datasets (NTU RGB+D [27], NTU RGB+D 120 [17] and UTKinect-Action3D Dataset [33]) with standard evaluation protocols. The results demonstrate the effectiveness of the proposed method for 3D skeleton-based action recognition.

In this paper, we introduce related work in Sect. 2. The process of selecting features from the original skeleton data and encoding them into grayscale images for CNN classification is described in Sect. 3. In Sect. 4, the implementation details of the experiments are explained, and the ablation study and experimental results are presented and compared with other related methods. Finally, Sect. 5 presents our conclusions.

## 2 Related Work

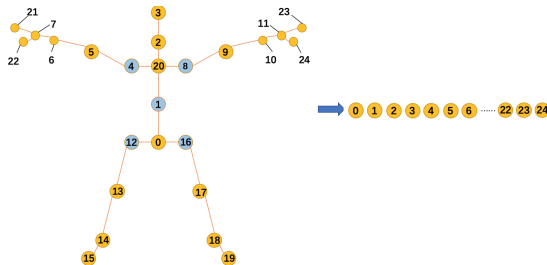
In this section, we introduce some related work on skeleton-based action recognition.

An end-to-end hierarchical RNN was proposed in [6]. The skeleton is divided according to the human body structure into five parts, which are input to five

networks for classification. In [5], the joint coordinates of a skeleton sequence are concatenated as a matrix, which is then quantified into an image for further classification, such that each image corresponds to a skeleton sequence. In [13], each skeleton sequence is used to generate four images, each of which incorporates the temporal information of the entire sequence and the spatial relationship in each frame. The four images incorporate different aspects of the spatial information of the whole sequence. To focus more attention on the most informative joints in skeleton frames, [22] proposed the global context-aware attention LSTM (GCA-LSTM). A two-stream RNN was proposed in [30] to handle the temporal and spatial information of a skeleton sequence. For temporal dynamics, a stacked RNN and a hierarchical RNN were proposed. For spatial information, a method of converting a spatial graph into a series of joints was proposed. In [19], a gating mechanism was used in an LSTM module to process the noise in a skeleton sequence. A method for multi-modal feature fusion using an LSTM was also proposed. Zhengyuan Yang et al. [34] proposed the idea of applying depth-first tree traversal in a skeleton representation, in the tree structure skeleton image (TSSI) method. Carlos Caetano et al. [1] combined the ideas of [13] and TSSI [34] to propose a new skeleton representation called the tree structure reference joints image (TSRJI). In the method of [24], features are separately extracted from the pose coordinate system and the trajectory coordinate system, which are the result of conversion from the original coordinate system of skeleton joints. These two types of features are then input to two LSTM networks and concatenated, for further classification.

### 3 The Proposed Method

This section introduces our method for generating the skeleton representation from original skeleton sequences. The skeleton sequences are trajectories of 3D coordinates of skeleton joints. The skeleton sequence is converted into several grayscale images, which incorporate both the dynamics and spatial information of the original skeleton structure. The method of calculating the relative distance between joints and important angles in the skeleton structure is explained.



**Fig. 1.** 3D skeleton joints structure. Joints 1, 4, 8, 12, 16 are key joints defined in [13]. Relative distance calculated by each key joint with other skeleton joints are preserved in an array according to the sequential order from 0 to 24.

The method of rearranging the pixels in the encoded images, depending on the spatial relationship between skeleton joints, can preserve much of the motion information. CNNs are then employed to extract features from these images for classification.

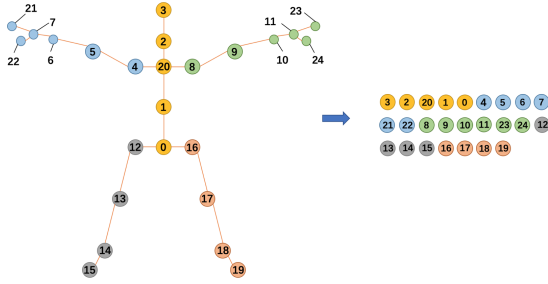
### 3.1 Encoding Process

In a skeleton sequence, each frame consists of several skeleton joints. The relative position of any two joints describes the motion migration. In previous work, the movement information of each frame has often been encoded to a single image; however, a long sequence contains too many images for this to be practical. According to [13], it is hard to learn temporal dynamics, and therefore each image will be very sparse. A new method for encoding the dynamics of skeleton sequences was proposed by [13] to overcome these weaknesses. In this paper, we propose an improved method using the skeleton representation of [13].

In the method of [13], the temporal dynamics of a skeleton sequence are encoded in an image, and the spatial information of the skeleton joints is incorporated into multiple images. It selects four key joints, which are considered to be stable throughout the action. But in our method, five key joints (the middle of the spine, left shoulder, right shoulder, left hip, and right hip) are chosen for use, as shown in Fig. 1. The relative distance between each key joint and other joints can then be computed. The skeleton joints in each skeleton frame are all numbered and arranged from 0 to 24. For each key joint, there is a corresponding array  $C$  with a size of  $N \times T$  ( $N$  is the number of skeleton joints in each frame and  $T$  is the number of frames in each sequence). Each value in the array is scaled to 0–255 by a linear transformation.

In the above encoding process for skeleton data, the arrangement of skeleton joints (i.e., the distance between a key joint and other joints in the calculated array) is in order of serial number, from 0 to 24, as shown in Fig. 1. However, in this order, the adjacent joints in each array lose some spatial relationships that are present in the original skeleton structures. For example, joint number 1 is the middle of the spine and joint number 20 is also the spine. They are adjacent in the original skeleton structure, but the direct adjacency relationship is lost after the encoding process. Conversely, joint number 11 is the right hand and joint number 12 is the left hip. Their numbers are consecutive, and so they are encoded adjacent to each other in the grayscale image, but they have no spatial relationship in the original skeleton structure. If this problem is not solved, the generated grayscale images are likely to lose many of the direct, or highly relevant, spatial relationships between the skeleton joints.

In this study, we propose a new method in which the skeleton joints are divided into groups according to the limb relationships. The positions of pixels in encoded images are changed according to that division, to enhance the spatial information preserved in encoded grayscale images. The original skeleton structure consists of 25 joints, which we divide into five groups according to the human body structure, as shown in Fig. 2. The five parts of the body are the



**Fig. 2.** Skeleton structure is divided into five parts: the trunk (3, 2, 20, 1, 0), the left hand (4, 5, 6, 7, 21, 22), the right hand (8, 9, 10, 11, 23, 24), the left leg (12, 13, 14, 15) and the right leg (16, 17, 18, 19). The new array is concatenated by the five groups as (3, 2, 20, 1, 0, 4, 5, 6, 7, 21, 22, 8, 9, 10, 11, 23, 24, 12, 13, 14, 15, 16, 17, 18, 19).



**Fig. 3.** The process of encoding skeleton representation from original skeleton sequence. For each input skeleton sequence, the order of the serial number of joints is first rearranged by our proposed method. Then the relative distance between key joint and other joints are calculated respectively. Each key joint corresponds to an encoded grayscale image and five images revealing the relative distance are generated.

trunk (3, 2, 20, 1, 0), left hand (4, 5, 6, 7, 21, 22), right hand (8, 9, 10, 11, 23, 24), left leg (12, 13, 14, 15), and right leg (16, 17, 18, 19). These five groups of joints are concatenated together to obtain a new array (3, 2, 20, 1, 0, 4, 5, 6, 7, 21, 22, 8, 9, 10, 11, 23, 24, 12, 13, 14, 15, 16, 17, 18, 19), as shown in Fig. 2. The five arrays, corresponding to the five key joints of a sequence, are used to generate five grayscale images. In this manner, the joints that are directly related in each part of the original skeleton structure are adjacent in the encoded images. The whole process of feature encoding is depicted in Fig. 3. The experiments in Sect. 4 indicate the effectiveness of our method of rearranging the order of joints according to the body joint relationships.

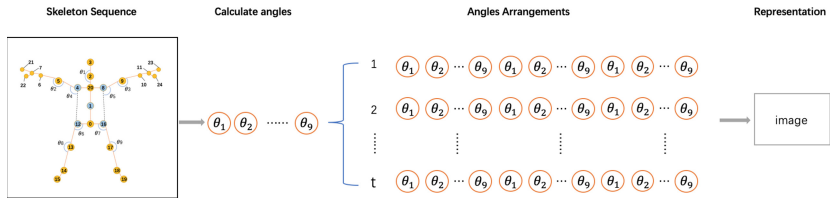
### 3.2 Important Angles Between Joints

Changes in the relative distance between joints in the skeleton structure can characterize action movements, so we encode five skeleton grayscale images, related to the five key joints from a skeleton sequence. Similarly, every pair of joints in a single skeleton frame can be treated as a vector, and the angle between two vectors can reveal the action gesture. In each skeleton sequence, a change of these angles provides motion information. In this paper, nine important angles are defined and calculated, as shown in Fig. 4. Each angle is defined by the following equation:

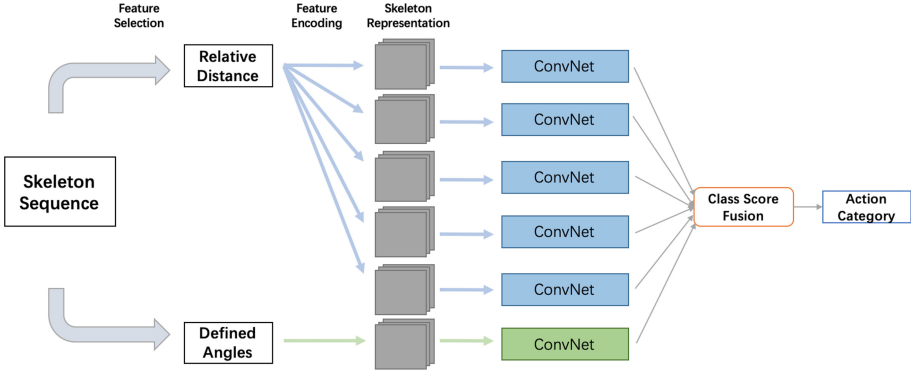
$$\theta = \cos^{-1}\left(\frac{\vec{a} \cdot \vec{b}}{|\vec{a}| \cdot |\vec{b}|}\right) \tag{1}$$

where  $\vec{a}$  and  $\vec{b}$  are vectors that include the same joint vertex and  $\theta$  is the angle between the two vectors. For example, as shown in Fig. 4,  $\theta_4$  is the angle between the arm and the trunk. To calculate  $\theta_4$ , three joints (4, 5, and 12) are selected, to form two 3D vectors using the joints' coordinates.  $\theta_4$  is calculated according to Eq. (1) and these vectors. For a skeleton sequence, we calculate the nine defined angles of each skeleton frame and arrange the values of each frame in a row. The values are quantified as an integer in  $[0, 255]$ , corresponding to the gray scale. The generation of the grayscale image for angle is similar to that of the grayscale image for relative distance, as shown in Fig. 4. However, in this section, the nine angle values are copied three times, to fill a row of the image, to enhance the effect of changes in angles. The effectiveness of this technique has been proved in our experiments. Thus, from each sequence, we generate a grayscale image that incorporates angle displacement information.

In summary, from each sequence, we generate five images by calculating the relative positions of joints and one image by computing important angles. These six grayscale images are then input to CNN for feature extraction and action classification. The integral pipeline is shown in Fig. 5.



**Fig. 4.** Nine angles labeled from  $\theta_1$  to  $\theta_9$  are defined in the left of the figure. These angles are calculated and quantified as an integer in  $[0, 255]$ , then the nine angle values are copied three times to fill each row of the image.



**Fig. 5.** The architecture of the recognition process. For each skeleton sequence, the relative distance corresponding to five key joints are computed to generate five skeleton images, and the defined angles are calculated to generate one skeleton image. The six grayscale images are fed into six identical CNNs for feature extraction simultaneously. At last, the class scores of different networks are fused to yield final classification results.

## 4 Experiments and Analysis

In this section, we evaluate the performance of our proposed multi-feature skeleton representation for 3D skeleton-based action recognition. The experiments were conducted on three 3D skeleton datasets: NTU RGB+D [27], NTU RGB+D 120 [17], and UTKinect-Action3D [33]. In particular, ablation experiments were conducted on the NTU RGB+D dataset, to assess the effectiveness of each component of our proposed model. We then compare our results with other methods on the three datasets, with standard evaluation protocols. The neural network architecture and implementation details are also explained.

### 4.1 Datasets

**NTU RGB+D Dataset [27].** This is a large dataset for action recognition with more than 56,000 video samples and 4 million frames, performed by 40 distinct subjects. It contains 60 action classes, including 40 daily actions (e.g., standing up, reading, and making a phone call), nine health-related actions (e.g., staggering and falling down), and 11 mutual actions (e.g., pushing, hugging, and shaking hands). All the data were collected by three Microsoft Kinect v2 cameras from different viewpoints simultaneously. This dataset provides four major data modalities: depth maps, RGB frames, IR sequences, and 3D skeleton sequences, consisting of 3D coordinates of 25 human body joints. Because of the large quantity and wide variety of views and classes, it is a challenging dataset.

**NTU RGB+D 120 Dataset [17].** This is the latest large-scale dataset for 3D skeleton-based human action recognition. It contains 120 action classes and more than 114,000 video clips, performed by 106 distinct subjects. It extends the NTU



RGB+D dataset [27] by adding more action classes and samples. It also provides daily, mutual, and health-related activities and several data modalities: depth maps, RGB frames, IR sequences, and 3D skeleton sequences. This dataset is very challenging because of its distinct action categories, varied human subjects, and diverse environmental conditions.

**UTKinect-Action3D Dataset** [33]. This dataset contains 10 categories of action classes: walk, sit down, stand up, pick up, carry, throw, push, pull, wave, and clap hands. These actions are performed by 10 subjects, each performing every action twice. There are 200 action clips and 6220 frames in total. The dataset is challenging because of its intra-class variation.

## 4.2 Implementation Details

The neural network architecture employed in our method is derived from temporal segment networks, which were proposed by [31], choosing Inception with batch normalization (BN-Inception) [11] as building block. Besides, the mean and variance parameter of Batch Normalization layers are frozen except the first one, and a dropout layer is added after the global pooling layer to reduce over-fitting. Each grayscale image is generated with the size of  $340 \times 256$ , the number of segments in the network is set to six, corresponding to our six generated grayscale images, which are input to six identical CNNs. The class scores of each image are fused by a function to obtain a consensus of class hypothesis among them, then the Softmax function is used to predict the action class based on the consensus, as [31] did.

In experiments on the NTU RGB+D and NTU RGB+D 120 datasets, we exploited models pretrained on the ImageNet [3] dataset, to initialize the network weights, and we use stochastic gradient descent to learn network parameters, as [31] did. We set the initial learning rate to 0.001, which is reduced by a factor of 10 after 10,000, 18,000, and 26,000 iterations. The maximum number of iterations was 36,000 and the batch size was set to 16. In experiments on the UTKinect-Action3D dataset, the model trained on the NTU RGB+D dataset by our method was used as the pretrained model, because of the small size of the UTKinect-Action3D dataset. The initial learning rate was 0.001, the stepsize was 1500, and the maximum number of iterations was 3500. The number of units in the fully connected layer was the same as the number of action categories in each dataset.

## 4.3 Experimental Evaluation

In this section, we show the results of the ablation evaluation on the NTU RGB+D dataset. We then present the recognition results on the three 3D skeleton-based datasets and make comparisons with other methods, to demonstrate the effectiveness of the proposed multi-feature skeleton representation.

**NTU RGB+D Dataset** [27]. The author of the dataset provided two protocols for training and testing. Cross-subject evaluation splits the dataset into

40,320 samples performed by 20 subjects for training, and 16,560 samples by the remaining 20 subjects for testing. Cross-view evaluation picks 37,920 samples from cameras 2 and 3 for training and 18,960 samples from camera 1 for testing. We followed the two protocols and report the experimental results from each.

**Table 1.** Ablation study and recognition performance compared with other methods on NTU RGB+D dataset. Accuracy on standard cross-subject and cross-view protocols are reported.

Method	Cross-subject	Cross-view
Dynamic skeletons [9]	60.23%	65.22%
Deep LSTM [27]	60.69%	67.29%
Part-aware LSTM [27]	62.93%	70.27%
ST-LASTM [20]	69.2%	77.7%
Two-stream RNN [30]	71.3%	79.5%
STA-LSTM [28]	73.4%	81.2%
TSRJI (late fusion) [1]	73.3%	80.3%
5 Key Joints	74.93%	79.11%
5 Key Joints + Angles	75.59%	80.56%
5 Key Joints + Rearrangement	76.03%	81.21%
5 Key Joints + Rearrangement + Angles	<b>77.33%</b>	<b>82.48%</b>

First, we examine the contributions of each component of our proposed representation, as shown in Table 1. The method of generating the skeleton representation named ‘5 Key Joints’ is based on the idea of [13], but the neural network architecture and implementation details are replaced. By reclassifying the skeletal structure and rearranging the pixels in the generated grayscale images according to the joints’ relevance, we can observe that ‘5 Key Joints + Rearrangement’ outperformed ‘5 Key Joints’ by 1.1% and 2.1% on the cross-subject and cross-view protocols, respectively. The results demonstrate the effectiveness of the proposed method of rearranging the order of joints according to the five body parts in the encoding process. It also indicates that two adjacent joints (connected by an edge) in the original skeleton structure incorporate more spatial information than two joints that are not adjacent. Our method allows the encoded images to preserve more spatial structure information of the skeleton joints that are adjacent to each other in space. Another component is the nine defined angles between joints in the skeleton structure; the calculated angles were added to ‘5 Key Joints’. ‘5 Key Joints + Angles’ achieved an accuracy of 75.59% and 80.56% on the cross-subject and cross-view protocols, respectively, outperforming ‘5 Key Joints’ on both protocols. This indicates that these angles play an important role in skeleton movements, so that the change of the angles reflects the action. This proves the effectiveness of the selected angles for action recognition. Finally, we combined the two methods (Rearrangement and Angles)

together. ‘5 Key Joints + Rearrangement + Angles’ achieved the best result, with 77.33% and 82.48% on the cross-subject and cross-view protocols, respectively. In particular, this result was obtained by using the strategy of copying the values of the angles three times as much as before, in the encoded grayscale image. We also conducted the experiments with no copying operation; the accuracy was 81.27% on the cross-view protocol and 76.17% on the cross-subject protocol, which is a worse result than that obtained with the copying operation. This indicates the effectiveness of the proposed strategy. Based on these results, we used this combined representation, ‘5 KeyJoints + Rearrangement + Angles’, in the following experiments for comparison with other methods.

We compared the proposed method with other methods using the NTU TGB+D dataset, as shown in Table 1. Our method achieved better results than others, on both the cross-subject and cross-view evaluation protocols. The best accuracy on the cross-view protocol was 82.48%, which is 2.18% higher than the best CNN-based method TSRJI (Late Fusion) [1], and higher than the best RNN-based method STA-LSTM [28] by 1.28%. On the cross-subject protocol, the proposed method achieved an accuracy of 77.33%. In comparison with the previous best method STA-LSTM [28], tested on the cross-subject protocol, the accuracy has been improved by 3.93%.

**Table 2.** Evaluation results on NTU RGB+D 120 Dataset. We list the accuracy on standard cross-subject and cross-setup protocols. Results of other methods for comparison are from [17].

Method	Cross-subject	Cross-setup
Dynamic skeletons [9]	50.8%	54.7%
Spatio-temporal LSTM [20]	55.7%	57.9%
Internal feature fusion [19]	58.2%	60.9%
GCA-LSTM [22]	58.3%	59.2%
Multi-task learning network [13]	58.4%	57.9%
FSNet [18]	59.9%	62.4%
Skeleton visualization (single stream) [23]	60.3%	63.2%
Two-stream attention LSTM [21]	61.2%	63.3%
Multi-task CNN with RotClips [14]	62.2%	61.8%
Magnitude-orientation (TSA) [2]	62.9%	63.0%
TSRJI (late fusion) [1]	<b>65.5%</b>	59.7%
5 key joints + rearrangement + angles	65.44%	<b>65.15%</b>

**NTU RGB+D 120 Dataset** [17]. This dataset is more challenging than the NTU RGB+D dataset, because it has many more action categories, performers, and skeleton sequences. There are two standard evaluation protocols for testing on this dataset. Cross-subject evaluation splits the 106 human subjects

into training and testing sets by the subjects’ identifiers. Cross-setup evaluation selects samples with even setup identifiers for training and odd identifiers for testing.

According to the experimental results shown in Table 1, the proposed method ‘5 Key Joints + Rearrangement + Angles’ achieved the best performance on the NTU RGB+D dataset, so we tested this method on the NTU RGB+D 120 dataset. As shown in Table 2, the proposed method outperformed all other methods on the cross-setup protocol, with an accuracy of 65.15%. On the cross-subject protocol, the best result was achieved by [1] with an accuracy of 65.5%. However, our method still achieved a competitive result, with an accuracy of 65.44%.

**UTKinect-Action3D Dataset** [33]. In the evaluation on the UTKinect dataset, we used the cross-subject evaluation protocol, as [26] did. Half of the subjects in the dataset were used for training and the remaining subjects were used for testing. The evaluation result of our method is shown in Table 3. The results of the other methods in Table 3 were all evaluated with the same cross-subject protocol. Compared with other methods, our method achieved the highest accuracy, 97.0%.

**Table 3.** Experimental results on UTKinect-Action3D Dataset.

Method	Accuracy
JL-d [35]	95.96%
Lie group representation [26]	96.68%
Ensemble TS-LSTM v2 [16]	96.97%
HST-RNN [32]	96.97%
5 key joints + rearrangement + angles	<b>97.0%</b>

## 5 Conclusions

In this paper, we proposed a method of extracting features from an original skeleton sequence as a multi-feature skeleton representation, which is then input to CNNs for further action classification. To enable the pixels in the generated grayscale images to preserve more of the spatial information of the original skeleton structure, we proposed the strategy of dividing the skeleton joints into several groups in which joints are related to each other, and rearranging the positions of pixels according to the relationships. This method proved effective for improving the recognition accuracy. In addition, a set of important angles in the skeleton structure was calculated, to form a multi-feature representation and utilize more spatial information. The proposed method was evaluated with standard evaluation protocols on three 3D skeleton-based action datasets: NTU RGB+D [27], NTU RGB+D 120 [17], and UTKinect-Action3D Dataset [33]. The experimental results indicated the effectiveness of our method of using a novel spatio-temporal skeleton representation for 3D skeleton-based action recognition.

**Acknowledgment.** The research in this paper used the NTU RGB+D and NTU RGB+D 120 Action Recognition Dataset made available by the ROSE Lab at the Nanyang Technological University, Singapore.

## References

1. Caetano, C., Brémond, F., Schwartz, W.R.: Skeleton image representation for 3D action recognition based on tree structure and reference joints. In: 2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), pp. 16–23. IEEE (2019)
2. Caetano, C., Sena, J., Brémond, F., Dos Santos, J.A., Schwartz, W.R.: Skelemotion: a new representation of skeleton joint sequences based on motion information for 3D action recognition. In: 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–8. IEEE (2019)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
4. Ding, Z., Wang, P., Ogunbona, P.O., Li, W.: Investigation of different skeleton features for CNN-based 3D action recognition. In: 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), pp. 617–622. IEEE (2017)
5. Du, Y., Fu, Y., Wang, L.: Skeleton based action recognition with convolutional neural network. In: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), pp. 579–583. IEEE (2015)
6. Du, Y., Wang, W., Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1110–1118 (2015)
7. Graves, A.: Supervised sequence labelling. In: Graves, A. (ed.) Supervised Sequence Labelling with Recurrent Neural Networks, pp. 5–13. Springer, Heidelberg (2012). [https://doi.org/10.1007/978-3-642-24797-2\\_2](https://doi.org/10.1007/978-3-642-24797-2_2)
8. Hbali, Y., Hbali, S., Ballihi, L., Sadgal, M.: Skeleton-based human activity recognition for elderly monitoring systems. *IET Comput. Vision* **12**(1), 16–26 (2017)
9. Hu, J.F., Zheng, W.S., Lai, J., Zhang, J.: Jointly learning heterogeneous features for RGB-D activity recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5344–5352 (2015)
10. Huang, C.D., Wang, C.Y., Wang, J.C.: Human action recognition system for elderly and children care using three stream convnet. In: 2015 International Conference on Orange Technologies (ICOT), pp. 5–9. IEEE (2015)
11. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv preprint [arXiv:1502.03167](https://arxiv.org/abs/1502.03167) (2015)
12. Kalayeh, M.M., Basaran, E., Gökmen, M., Kamasak, M.E., Shah, M.: Human semantic parsing for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1062–1071 (2018)
13. Ke, Q., Bennamoun, M., An, S., Sohel, F., Boussaid, F.: A new representation of skeleton sequences for 3D action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3288–3297 (2017)
14. Ke, Q., Bennamoun, M., An, S., Sohel, F., Boussaid, F.: Learning clip representations for skeleton-based 3D action recognition. *IEEE Trans. Image Process.* **27**(6), 2842–2855 (2018)
15. LeCun, Y., Bengio, Y., et al.: Convolutional networks for images, speech, and time series. *Handb. Brain Theory Neural Netw.* **3361**(10), 1995 (1995)

16. Lee, I., Kim, D., Kang, S., Lee, S.: Ensemble deep learning for skeleton-based action recognition using temporal sliding LSTM networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1012–1020 (2017)
17. Liu, J., Shahroudy, A., Perez, M.L., Wang, G., Duan, L.Y., Chichung, A.K.: NTU RGB+D 120: a large-scale benchmark for 3D human activity understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* **42**(10), 2684–2701 (2019)
18. Liu, J., Shahroudy, A., Wang, G., Duan, L.Y., Chichung, A.K.: Skeleton-based online action prediction using scale selection network. *IEEE Trans. Pattern Anal. Mach. Intell.* **42**(6), 1453–1467 (2019)
19. Liu, J., Shahroudy, A., Xu, D., Kot, A.C., Wang, G.: Skeleton-based action recognition using spatio-temporal LSTM network with trust gates. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**(12), 3007–3021 (2017)
20. Liu, J., Shahroudy, A., Xu, D., Wang, G.: Spatio-temporal LSTM with trust gates for 3D human action recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) *ECCV 2016. LNCS*, vol. 9907, pp. 816–833. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-46487-9\\_50](https://doi.org/10.1007/978-3-319-46487-9_50)
21. Liu, J., Wang, G., Duan, L.Y., Abdiyeva, K., Kot, A.C.: Skeleton-based human action recognition with global context-aware attention LSTM networks. *IEEE Trans. Image Process.* **27**(4), 1586–1599 (2017)
22. Liu, J., Wang, G., Hu, P., Duan, L.Y., Kot, A.C.: Global context-aware attention LSTM networks for 3D action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1647–1656 (2017)
23. Liu, M., Liu, H., Chen, C.: Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recogn.* **68**, 346–362 (2017)
24. Pan, G., Song, Y., Wei, S.: Combining pose and trajectory for skeleton based action recognition using two-stream RNN. In: 2019 Chinese Automation Congress (CAC), pp. 4375–4380. IEEE (2019)
25. Pan, H., Chen, Y.: Multilevel LSTM for action recognition based on skeleton sequence. In: 2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), pp. 2218–2223. IEEE (2019)
26. Rhif, M., Wannous, H., Farah, I.R.: Action recognition from 3D skeleton sequences using deep networks on lie group features. In: 2018 24th International Conference on Pattern Recognition (ICPR), pp. 3427–3432. IEEE (2018)
27. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: NTU RGB+D: a large scale dataset for 3D human activity analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1010–1019 (2016)
28. Song, S., Lan, C., Xing, J., Zeng, W., Liu, J.: Spatio-temporal attention-based LSTM networks for 3D action recognition and detection. *IEEE Trans. Image Process.* **27**(7), 3459–3471 (2018)
29. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: person retrieval with refined part pooling. In: *ECCV* (2018)
30. Wang, H., Wang, L.: Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 499–508 (2017)
31. Wang, L., et al.: Temporal segment networks: towards good practices for deep action recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) *ECCV 2016. LNCS*, vol. 9912, pp. 20–36. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-46484-8\\_2](https://doi.org/10.1007/978-3-319-46484-8_2)

32. Wei, S., Song, Y., Zhang, Y.: Human skeleton tree recurrent neural network with joint relative motion feature for skeleton based action recognition. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 91–95. IEEE (2017)
33. Xia, L., Chen, C., Aggarwal, J.: View invariant human action recognition using histograms of 3D joints. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 20–27. IEEE (2012)
34. Yang, Z., Li, Y., Yang, J., Luo, J.: Action recognition with spatio-temporal visual attention on skeleton image sequences. *IEEE Trans. Circuits Syst. Video Technol.* **29**(8), 2405–2415 (2018)
35. Zhang, S., Liu, X., Xiao, J.: On geometric features for skeleton-based action recognition using multilayer LSTM networks. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 148–157. IEEE (2017)
36. Zheng, Z., Zheng, L., Yang, Y.: Pedestrian alignment network for large-scale person re-identification. *IEEE Trans. Circuits Syst. Video Technol.* **29**(10), 3037–3045 (2018)