Three-Stream Convolutional Neural Network with Multi-task and Ensemble Learning for 3D Action Recognition

Duohan Liang[1*], Guoliang Fan[1,2], Guangfeng Lin[1], Wanjun Chen[1], Xiaorong Pan[1+], Hong Zhu[1]

[1]Xi'an University of Technology, Xi'an, Shaanxi 710048, China
[2]Oklahoma State University, Stillwater, OK 74078, USA

{2170320179*, 2180320181+}@stu.xaut.edu.cn
{guoliang.fan, lgf78103, wjchen, zhuhong}@xaut.edu.cn

Abstract

In this paper, we propose a three-stream convolutional neural network (3SCNN) for action recognition from skeleton sequences, which aims to thoroughly and fully exploit the skeleton data by extracting, learning, fusing and inferring multiple motion-related features, including 3D joint positions and joint displacements across adjacent frames as well as oriented bone segments. The proposed 3SCNN involves three sequential stages. The first stage enriches three independently extracted features by co-occurrence feature learning. The second stage involves multi-channel pairwise fusion to take advantage of the complementary and diverse nature among three features. The third stage is a multi-task and ensemble learning network to further improve the generalization ability of 3SCNN. Experimental results on the standard dataset show the effectiveness of our proposed multi-stream feature learning, fusion and inference method for skeleton-based 3D action recognition.

1. Introduction

Human action recognition has been an active topic in computer vision, as it has a wide range of applications such as video understanding, intelligent surveillance system, human-computer interaction and so on [8, 12, 14, 17]. Compared with the conventional RGB videos recorded by two-dimensional cameras, the skeleton-based sequences contain three-dimensional (3D) coordinates of key joints of the human body, which can provide an effective and robust representation for describing human actions with complicated background [8, 11, 19]. Thus, representation methods based on skeleton data have attracted considerable attention for action recognition in recent years.

Considering the time correlation of actions in skeletal videos, many of the early works regard recurrent neural networks (RNN) as natural choices [3, 15, 17]. Although the RNN-based models were designed to model the temporal dependency, it is difficult to train the stacked RNN so that the high-level feature cannot be learned directly from skeletons [4, 11]. CNN models can easily to build the deep network and equipped with excellent ability to extract high-level information. Thus, many researchers encode the skeleton joints to multiple 2D pseudo-images, and feed them into a multi-stream framework with CNN to separately learn spatial-temporal features, and then perform a simple feature fusion [1, 6, 7, 10, 18, 19]. Despite the significant improvements in performance, there exist two problems to be addressed. First, features used in the above methods merely focus on the displacement of skeleton joints in time and space. The body's size, direction and other attributes play an important role in skeleton action recognition, however, these features are usually neglected. Second, most work-

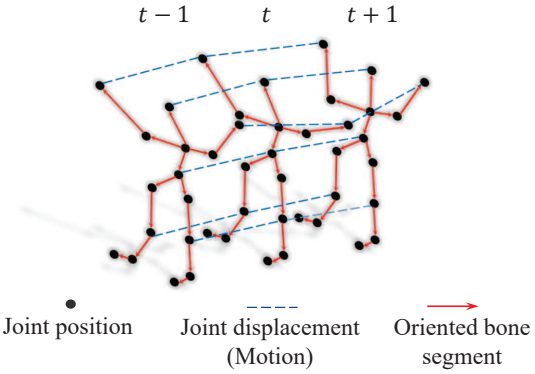

Figure 1. The illstration of three motion-related features extracted from skeleton sequences, including 3D joint positions (denoted by black points), joint displacements across adjacent frames (i.e., motion vectors shown by blue dashed lines) and oriented bone segments (represented by the red vectors).

s based on a multi-stream framework regard each stream as independent in the feature extraction stage and take a simple feature fusion in the final classification stage [1, 7, 17–19], which may not effectively exploit the complementarity and diversity among multiple features.

To address these issues, we propose a novel three-stream CNN (3SCNN) model which can comprehensively exploit information of skeleton sequences. Firstly, we design the feature of bone segment containing information of body's length and direction, and combine bone segment with skeleton position and skeleton motion [10, 11], as shown in Fig 1, for a comprehensive description of skeleton-based action. Thus, we use a three-stream framework to handle the three features. Secondly, we propose pairwise feature fusion and a multi-task and ensemble learning network for better fusing information from multiple features in different stages and exploring the relationship among multiple features. Fig 2 shows the flowchart of our model that contains three stages. In the first stage, skeleton's position, motion and bone segment after the view adaption processing extract their own global co-occurrence feature [11] independently. In the second stage, the pairwise feature fusion is performed on the output of the previous stage. The third stage is the multi-task learning network during training and is the ensemble learning network during inference for better performance on action recognition.

The main contributions of this paper are listed as follows:

- We introduce a new feature that is the oriented bone segment from the perspective of the subject in action. The bone segments are combined with skeleton position and the skeleton motion in a three-stream framework for improving the recognition performance.
- We propose pairwise feature fusion to fully take advantage of the complementary and diverse nature among three features by a two-stage fusion strategy.
- We design a multi-task and ensemble learning network to further improve generalization ability of the model and our 3SCNN obtains the state-of-the-art results on the largest in-door dataset NTU RGB+D.

2. Related Work

We briefly review the existing literature in related research from two perspectives as follows.

RNN-based and CNN-based approaches for skeleton-based action recognition RNN-based network has become prevalent due to its advantage for modeling sequence data. Part-aware LSTM [15] designed a part aware extension of LSTM to take advantage of the physical structure of the human body. The work of [22] introduced a regularization scheme to a fully connected deep LSTM network in order to automatically learn co-occurrence feature. View-Adaptive RNN (VA-RNN) [20] designed a view adaptive subnetwork in LSTM-based model, which assists the network in selecting the most suitable virtual observation viewpoint. Ensemble temporal sliding LSTM (TS-LSTM) [5] utilized an average ensemble to merge multiple parts containing short-term, medium-term, long-term temporal dependencies and even spatial skeleton pose dependency. Two-stream RNN [17] architecture was employed to separately model both spatial and temporal relations of joints of skeleton and then leveraged the features from each stream to weighted fuse. Compared with RNN, there is a growing tendency of using CNN for skeleton-based action recognition owing to its good parallel ability and easier training process. Skeleton-based CNN (SK-CNN) [2] treated skeleton sequence as 2D pseudo image then employed CNNs for skeleton-based action recognition. Ke *et al.* [4] proposed a new representation of skeleton sequences based on cylindrical coordinate and fed the new representation to a multi-task learning deep convolutional neural network for action recognition. According to the property of skeleton sequences, ensemble neural network (Ensem-NN) [19] designed four different subnets and fused them using one ensemble network. Two-stream CNN [10] employed a two-stream framework to combine position and motion information of human joints. Then motivated by co-occurrence learning [22], hierarchical co-occurrence network (HCN) [11] was proposed which utilized CNN to learn co-occurrence and achieve state-of-the-art performance.

Multiple feature learning and feature fusion For comprehensively describing skeleton sequence, multiple features usually are combined to represent skeleton action. Wang *et al.* [18] proposed Joint Trajectory Maps(JTM) to encode body joint trajectories and employed multiply-score fusion to the three JTMs projected onto three Cartesian planes. Li *et al.* [7] encoded the joint distances in the three orthogonal 2D planes and in the 3D space into four joint distance maps, and combined information of four spaces for action recognition. Liu *et al.* [13] leveraged multi-stream CNN to learn fusion feature from 10 kinds of spatial-temporal images with skeleton encoding. Li *et al.* [6] found different size skeleton images encoded by the translation-scale invariant image mapping method bring different frequency variance, thus adopted the multi-scale CNN model and the average fusion strategy to combine the multi-frequency information.

Inspired by the above works, we adopt multi-stream CNN to extract features from the skeleton's position, motion and bone segment. In addition, we design the pairwise feature fusion for comprehensively using information that is implicit in different features and introduce the idea of multi-task learning and ensemble learning to our 3SCNN for improving the generalization ability of the model.

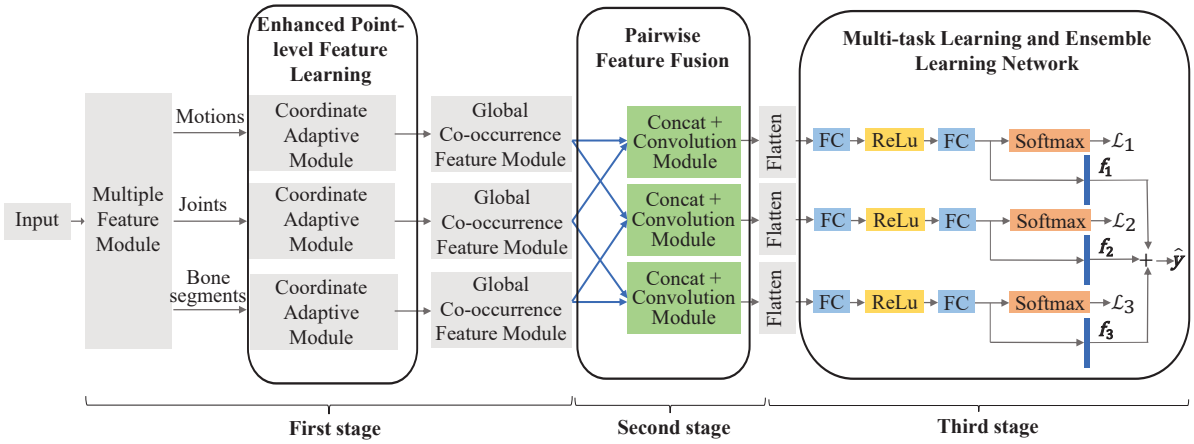

Figure 2. The flowchart of 3SCNN. The first stage independently extracts features. The second stage is feature fusion. The third stage is a multi-task learning network during training and is an ensemble learning network during inference.

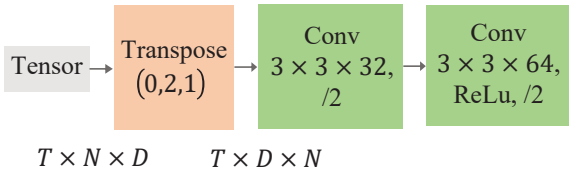

Figure 3. The architecture of the global co-occurrence feature learning module. $T \times N \times D$ is the input tensor's size. $T \times D \times N$ is tensor's size after the transpose layer processing. Transpose layer permutes the dimensions of the input tensor according to the order parameter. Conv represent the convolution layer, in which the last dimension denotes the number of output channels. The ReLu is activation function. The "/2" stand for an appended Maxpooling layer with stride 2 after convolution. The following figure annotations use the same mark.

3. Proposed Three-stream CNN Framework

In this section, we first briefly review the co-occurrence feature learning with convolutional neural network. Then we describe three stages of our proposed three-stream CNN (3SCNN) separately.

3.1. Co-occurrence Feature Learning with CNN

An action is usually only associated with and characterized by the interactions and combinations of a subset of skeleton joints [22]. For example, for the action of "making a phone call", the joints "hand", "arm", and "head" constitute the discriminative set of joints. Some joints set of the skeleton is considered as a co-occurrence feature that can intrinsically characterize a human action.

In HCN [11], the convolution operation is decomposed into two steps. In the first step, an independent convolution kernel slides on each channel of input so that the features are aggregated locally from neighborhoods of kernel. In the second step, an element-wise summation across channel is used for global features aggregation from all channels of inputs. So, Li *et al.* [11] suggest that the information if it is specified as channels, can be aggregated globally.

In traditional CNN-based methods [6, 10], they cast the frame, joint and coordinate dimensions of a skeleton sequence into width, height and channel of an image respectively. It causes the co-occurrence features to only be aggregated locally. For globally aggregating co-occurrence features, HCN corresponds to the joint dimension to channels by transposing the tensor of CNN's input. In addition, HCN can step by step learn the point-level representation and global co-occurrence features. Following the main idea of HCN, we introduce the point-level convolution and global co-occurrence feature learning module for which the architecture is shown in Fig 3, to our model.

3.2. Feature Enhancement

In order to enrich expression of skeleton sequences, we introduce the bone segment feature and multi-coordinate transformation to the point-level feature learning stage.

Multi-Feature Module Existing methods based on two-stream architecture leverage coordinates and temporal movements of joints as input for action recognition. Besides these information, bone segments between adjacent joints also provide the crucial cues to describe the human action because bone segments can directly reflect the body's length and direction information. We explicitly propose a model to regard the bone segment feature as another stream combined with two-stream architecture, and construct the three-stream network. We define $J = (x, y, z)^T$ that is a

3D joint coordinate. In frame t, S^T is raw skeleton coordinates, M^T describe raw skeleton motion, and B^T stands for the bones information. They are formulated as:

$$S^T = \left\{J_1^t, J_2^t, \cdots, J_N^t\right\}, \tag{1}$$

$$M^T = \left\{J_1^{t+1} - J_1^t, J_2^{t+1} - J_2^t, \cdots, J_N^{t+1} - J_N^t\right\}, \tag{2}$$

$$B^T = \left\{B_i^T = J_n^{t+1} - J_m^t, i = 1, 2, \cdots, N-1\right\}, \tag{3}$$

where N is the number of joints. n and m are the index of adjacent two joints. So, the number of bone segments equals $N-1$.

Coordinate-Adaptive Module The same action captured by different camera viewpoints provides the variously discriminative information. Intuitively, if the various skeleton sequences information at multiple coordinate systems are combined, the action will gain more comprehensive expression. A skeleton sequence at arbitrary viewpoint can be attained by rotation in 3D space. Therefore, given a 3×3 rotation matrix R_i, some rotations of one skeleton sequence $\varsigma = \{S^t, t = 1, 2, \cdots, \mathrm{T}\}$ can be represented as:

$$\varsigma_i = (\varsigma R_i)^T = \left\{S^1 R_i, S^2 R_i, \cdots, S^{\mathrm{T}} R_i\right\}^T, \tag{4}$$

where ς_i is the sequence after rotation. For exploiting the completer information from different aspects, we employ $R_1, R_2, \cdots, R_L$ rotation matrix to transform original ς corresponding to obtain L new sequences. These rotated sequences are concatenated as the set $\{\varsigma_1, \varsigma_2, \cdots, \varsigma_L\}$. The above operations are called multi-coordinate transformation. Then, we perform point-level convolution operation which consists of 1×1 convolution layer that adaptively combines sequences of multiple coordinate systems and 1×3 convolution layer that extracts the point-level feature in temporal. These rotation matrix R_i learn from the skeleton data, so we achieve the operation with L fully connected layers. Fig 4 shows the details of coordinate-adaptive module architecture which consists of multi-coordinate transformation and point-level convolution.

3.3. Pairwise Feature Fusion Learning

In the first stage, position, motion and bone segment features are extracted independently. To effectively exploit the complementarity and diversity among the three features, we proposed Pairwise Feature Fusion (PFF). As illustrated in Figure 5, PFF consists of two procedures which are pairing and fusing for each feature. In pairing, any two of three features can be made a pair by concatenating operation. Then, there are two alternative fusion architectures (shared fusion and split fusion) in fusing. In *split fusion*, each pairwise feature possesses exclusive fusion block to learn fusion pattern

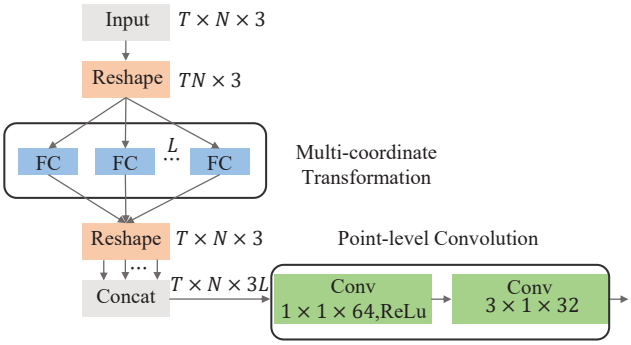

Figure 4. The architecture of coordinate-adaptive module. The input is raw data where T is number of frames, N is number of joints. Reshape layer transforms the input tensor to a certain shape. Concat layer concatenate the various same shape tensor on channel dimension. L is the number of fully connected layers (FC). $L = 10$ in our experiments.

respectively. In *shared fusion*, learning the fusion pattern of the three pairs uses one shared block. For those activities involving human-human interaction, we follow Li *et al.* [10, 11] adopting element-wise max scheme for the features of multiple person at end of pairwise feature learning stage.

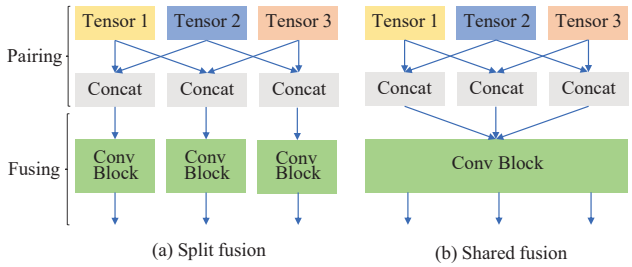

Figure 5. The two kinds of architecture of pairwise feature fusion. Conv block contains two convolution layers with kernel size of 3×3 and channels of 128, 256 respectively, and ReLu activation function and maxing-pooling are applied on each layer.

3.4. Multi-task and Ensemble Learning Network

As shown in the third stage in Fig. 2, three features after pairwise fusion learning are sent to their own classifier. Consequently, our three-stream model predicts three probability vectors $\boldsymbol{p}^1, \boldsymbol{p}^2, \boldsymbol{p}^3$ for each action. During training, we optimize it as a multi-task learning problem with cross-entropy loss and each classifier produces a loss component as:

$$\mathcal{L}_k = -\sum_{i=1}^{c} \boldsymbol{y}_i \log\left(\boldsymbol{p}_i^k\right), \tag{5}$$

where $k \in \{1, 2, 3\}$ is the number of each stream, $\boldsymbol{y}$ is the one hot vector of true label, and c is the number of action categories. Thus, the final loss can be defined as:

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3. \tag{6}$$

During inference, we refer to the main idea of ensemble learning, which is a machine learning paradigm with multiple learners trained for the same task, for better performance.The output $\boldsymbol{f}_i$ of the last fully connected layer of each classifier is jointly used to make a decision for human action recognition. For mitigating the high level of noise to make ambiguous classification, we choose the sum rule to joint these feature vectors. It can be represented as follows:

$$\hat{\boldsymbol{y}} = \text{softmax}\left(\sum_{i=1}^{k} \boldsymbol{f}_i\right), \tag{7}$$

where $\hat{\boldsymbol{y}}$ is the vector of final predicted class probabilities. Assuming the posterior probability is $P(j|\hat{\boldsymbol{y}})$ for class j, the final category belonging to class c can computed by

$$c = \underset{j}{\text{argmax}} P(j|\hat{\boldsymbol{y}}). \tag{8}$$

4. Experiments

We verify the effectiveness of our proposed model on a common benchmark dataset, the NTU RGB+D [15]. To find the impact of each component in our model, we perform the ablation study on the dataset, and we also compare and analyze the results of different fusion structures.

4.1. Datasets and Implementation Details

NTU RGB+D To the best of our knowledge, the dataset is currently the largest in-door daily skeleton-based action recognition dataset, which contains more than 56000 sequences in 60 classes of action performed by 40 subjects. Each sequence consists of 25 joints with one or two persons. The large intra-class and view point variations make the datasets have two recommended evaluation protocols, i.e. Cross-Subject (CS) and Cross-View (CV). For the cross-subject setting, the sequences of 20 subjects are used for training and the rest from other 20 subjects are used for testing. For the cross-view setting, samples are split by camera views, where two view-points are used for training and the rest for testing.

During the data processing, we randomly crop subsequence from entire sequence for data augment. The cropping ratio is drawn from uniform distribution between [0.5,1] for training and then is fixed with 0.9 for inference. Due to the variety in action length, we normalize the sequence to a fixed length 32 with bilinear interpolation along the frame dimension.

During training, we apply the weight decay of 0.001 on the weights of the first fully connected layer of each classifier. The network is trained using the Adam optimizer. The learning rate is initialized to 0.001, followed by an exponential decay with a rate of 0.99 per 1k batches, and the batch size is set to 64. The training is stopped when the learning inclines to 0.00001. We append the dropout with ratio of 0.5 on last convolution layer of the global feature learning module of each stream, on all layers in pairwise feature fusion and on the first fully connected layer of each classifier.

4.2. Ablation Study

Importance of multi-task and ensemble learning network To understand the impact of the third stage network in our model, we perform an ablation study. We deliberately use element-wise sum to merge the three classifiers' output to one and then only use one loss to optimize the model which is shown in Fig 6 (a). The modified network is referred to as 3SCNN-1L without multi-task learning and ensemble learning. We train 3SCNN-1L model with the same hyper-parameters as 3SCNN. The results are listed in Table 1. Compared with 3SCNN-1L, the multi-task learning network can make output of each stream maximize its role as much as possible. To further understand the effect of ensemble learning network in 3SCNN model, we also list the three classifiers' result, which are 3SCNN-1S, 3SCNN-2S and 3SCNN-3S shown in Fig 6 (b), before ensemble learning in Table 1. Table 1 shows that combining all classifier predicts using ensemble learning network will attain stronger generalization ability than each independent predict. In the following experiments the multi-task learning and ensemble learning network is adopted.

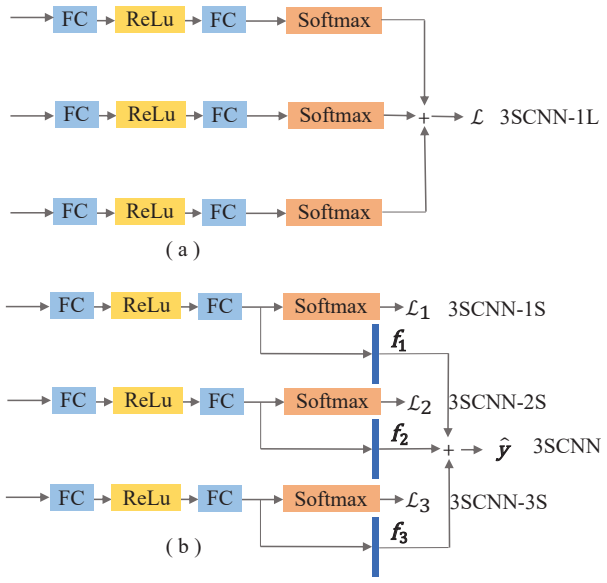

Figure 6. The two different third stages of our proposed model.

Methods	CS	CV
3SCNN-1L	87.1	92.5
3SCNN-1S	86.4	92.2
3SCNN-2S	87.5	92.5
3SCNN-3S	83.9	89.5
3SCNN	88.6	93.7

Table 1. The cross-view and cross-subject performance of the multi-task learning and ensemble learning network.

Comparisons on different fusion strategies Firstly, we evaluate the two fusion strategies proposed in section 3.3. As shown in Table 2, *split fusion* has higher precision than *shared fusion*, while *shared fusion* has fewer parameters with less precision loss. The reason might be that fusion between different features has its own unique fusion pattern and thus *split fusion* can better maintain the uniqueness in fusing than *shared fusion*. To further understand the behavior of pairwise feature fusion, we perform an ablation study. We replace pairwise feature fusion (PFF) with all concatenate fusion (ACF) which concatenate all features from previous stage in channel dimension and then feed it to one convolution block for extracting fusion feature, thus three streams are merged to one stream in here. The above operation is equivalent to removing the pairing process in PFF. So, the following network use only one classifier and the modified model is optimized with one loss. We can see that model with PFF outperforms the model with ACF. It indicates that PFF can better exploit relationships among features. To avoid the impact of multi-task and ensemble learning network, comparing 3SCNN-ACF and 3SCNN-1L further directly reflects the effectiveness of PFF.

Methods	CS	CV
3SCNN(PFF-Shared)	88.3	93.4
3SCNN(PFF-Split)	88.6	93.7
3SCNN-ACF	86.5	91.5
3SCNN-1L	87.1	92.5

Table 2. Evaluation of different fusion methods.

Impact of multi-coordinate transformation The core of coordinate-adaptive module described in section 3.2 is multi-coordinate transformation (MCT). Thus, we reduce our original 3SCNN method by removing the MCT step, leading to an algorithm referred to as 3SCNN*, where the raw joints, raw motions and raw bone segments are used as the inputs. The results in Table 3 demonstrate that the model with complete coordinate-adaptive module achieve a better performance. It indicates that MCT can provide more discriminative information from different coordinate space, and point-level convolution operation can effectively combine these information.

Methods	CS	CV
3SCNN* (without MCT)	88.2	93.3
3SCNN (with MCT)	88.6	93.7

Table 3. Evaluation of multi-coordinate transformation.

4.3. Comparison to Other State-of-the-art Methods

We compare the performance of our proposed model with several state-of-the-art methods on the NTU dataset in Table 4. Our method achieves the best action recognition accuracy in both CS and CV protocols. Compared with the RNN-based approach [21], the accuracy is improved by 7.9% in cross-subject setting and 5.3% in cross-view setting. And our 3SCNN outperforms the CNN-based method [11] by 2.1% in cross-subject setting and 2.6% in cross-view setting.

Methods	CS	CV	Year
Part-aware LSTM [15]	62.9	70.3	2016
Two-stream RNN [17]	71.3	79.5	2017
Ensemble TS-LSTM [5]	74.6	81.3	2017
VA-RNN [20]	79.4	87.6	2017
Clips + CNN + MTLN [4]	79.6	84.8	2017
LSTM + CNN [9]	82.9	90.1	2017
Two-stream CNN [10]	83.2	89.3	2017
Ensem-NN [19]	85.1	91.3	2018
HCN [11]	86.5	91.1	2018
EleAtt-GRU [21]	80.7	88.4	2018
SR-TSL [16]	84.8	92.4	2018
3SCNN	88.6	93.7	

Table 4. Comparisons of the proposed method with the previous approaches for action recognition on the NTU RGB+D dataset.

5. Conclusion

In this paper, we have proposed a novel three-stream CNN model, referred to as 3SCNN, for skeleton-based action recognition. To effectively exploit the joint, motion and bone segment features, we introduced a three-stream framework to handle the three kinds of inputs jointly. To further enrich feature expression, we designed the coordinate-adaptive module. We also proposed a pairwise feature fusion scheme and a multi-task ensemble learning network to take advantage of the complementary and diverse nature among multiple features. The proposed 3SCNN model shows impressive performance when compared with the state-of-the-art algorithms the NTU RGB+D dataset.

Acknowledgment

The authors would like to thank the reviewers for their comments that help improve the quality of this paper. This work was supported by the Shaanxi Hundred Talent Program and the NSFC Grant (Program No.61771386).

References

[1] Z. Ding, P. Wang, P. O. Ogunbona, and W. Li. Investigation of different skeleton features for cnn-based 3d action recognition. In *Proc. ICMEW*, pages 617–622, 2017.

[2] Y. Du, Y. Fu, and L. Wang. Skeleton based action recognition with convolutional neural network. In *Proc. ACPR*, pages 579–583, 2015.

[3] Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proc. CVPR*, pages 1110–1118, 2015.

[4] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid. A new representation of skeleton sequences for 3d action recognition. In *Proc. CVPR*, pages 3288–3297, 2017.

[5] I. Lee, D. Kim, S. Kang, and S. Lee. Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks. In *Proc. ICCV*, pages 1012–1020, 2017.

[6] B. Li, Y. Dai, X. Cheng, H. Chen, Y. Lin, and M. He. Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep cnn. In *Proc. ICMEW*, pages 601–604, 2017.

[7] C. Li, Y. Hou, P. Wang, and W. Li. Joint distance maps based action recognition with convolutional neural networks. *IEEE Signal Processing Letters*, 24:624–628, 2017.

[8] C. Li, S. Sun, X. Min, W. Lin, B. Nie, and X. Zhang. End-to-end learning of deep convolutional neural network for 3d human action recognition. In *Proc. ICMEW*, pages 609–612, 2017.

[9] C. Li, P. Wang, S. Wang, Y. Hou, and W. Li. Skeleton-based action recognition using lstm and cnn. In *Proc. ICMEW*, pages 585–590, 2017.

[10] C. Li, Q. Zhong, D. Xie, and S. Pu. Skeleton-based action recognition with convolutional neural networks. In *Proc. ICMEW*, pages 597–600, 2017.

[11] C. Li, Q. Zhong, D. Xie, and S. Pu. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. In *Proc. IJCAI*, pages 786–792, 2018.

[12] J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva, and A. C. Kot. Skeleton-based human action recognition with global context-aware attention lstm networks. *IEEE Transactions on Image Processing*, 27:1586–1599, 2018.

[13] M. Liu, H. Liu, and C. Chen. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, 68:346–362, 2017.

[14] L. L. Presti and M. La Cascia. 3d skeleton-based human action classification: A survey. *Pattern Recognition*, 53:130–147, 2016.

[15] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proc. CVPR*, pages 1010–1019, 2016.

[16] C. Si, Y. Jing, W. Wang, L. Wang, and T. Tan. Skeleton-based action recognition with spatial reasoning and temporal stack learning. In *Proc. ECCV*, pages 103–118, 2018.

[17] H. Wang and L. Wang. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In *Proc. CVPR*, pages 499–508, 2017.

[18] P. Wang, Z. Li, Y. Hou, and W. Li. Action recognition based on joint trajectory maps using convolutional neural networks. In *Proc. ACM MM*, pages 102–106, 2016.

[19] Y. Xu, J. Cheng, L. Wang, H. Xia, F. Liu, and D. Tao. Ensemble one-dimensional convolution neural networks for skeleton-based action recognition. *IEEE Signal Processing Letters*, 25:1044–1048, 2018.

[20] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In *Proc. ICCV*, pages 2117–2126, 2017.

[21] P. Zhang, J. Xue, C. Lan, W. Zeng, Z. Gao, and N. Zheng. Adding attentiveness to the neurons in recurrent neural networks. In *Proc. ECCV*, pages 135–151, 2018.

[22] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie. Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In *Proc. AAAI*, pages 3679–3703, 2016.