electronics

MDPI

Article

A Multi-Feature Representation of Skeleton Sequences for Human Interaction Recognition

Xiaohang Wang and Hongmin Deng *

School of Electronics and Information Engineering, Sichuan University, Chengdu 610065, Sichuan, China; xhwangg@stu.scu.edu.cn

* Correspondence: hm_deng@scu.edu.cn

Received: 24 November 2019; Accepted: 13 January 2020; Published: 19 January 2020

Abstract: Inspired from the promising performances achieved by recurrent neural networks (RNN) and convolutional neural networks (CNN) in action recognition based on skeleton, this paper presents a deep network structure which combines both CNN for classification and RNN to achieve attention mechanism for human interaction recognition. Specifically, the attention module in this structure is utilized to give various levels of attention to various frames by different weights, and the CNN is employed to extract the high-level spatial and temporal information of skeleton data. These two modules seamlessly form a single network architecture. In addition, to eliminate the impact of different locations and orientations, a coordinate transformation is conducted from the original coordinate system to the human-centric coordinate system. Furthermore, three different features are extracted from the skeleton data as the inputs of three subnetworks, respectively. Eventually, these subnetworks fed with different features are fused as an integrated network. The experimental result shows the validity of the proposed approach on two widely used human interaction datasets.

Keywords: convolutional neural networks; interaction recognition; skeleton joints

1. Introduction

Human action recognition and interaction recognition have recently attracted the intensive attention of researchers in computer vision field due to its extensive application prospects, such as intelligent surveillance, human-machine interaction, and so on. Most previous methods are devoted to the human action recognition in two-dimensional RGB data [1,2]. However, due to the high sensitivity to environmental variability of the RGB data, precise action recognition is also a challenging task. Some previous works have made some contributions to overcome these challenges [3,4]. In [3], Rezazadegan et al. proposed an action region proposal method that, informed by optical flow to extract image regions likely to contain actions, which can eliminate the influence of background. Besides, this problem could be overcome by using cost-efficient RGB-D (i.e., color plus depth) sensors [5]. Generally, depth sensors can provide three-dimensional (3D) information of human body in more detail and with high robustness to variations of perspectives [6]. Therefore, human action recognition based on 3D skeleton has become a hot research topic.

Human action can be described by a series of time sequences of skeleton. The temporal information and spatial information are significant for skeleton based action recognition. Several types of recurrent neural networks (RNN), including long short-term memory (LSTM) [7] and gated recurrent units (GRU) [8], showed great advantages on processing sequence data. However, the temporal modeling is not always suitable for the skeleton sequences. On the other hand, the RNN lacks the ability of spatial modeling, while convolutional neural networks (CNN) have natural advantages in extracting spatial features. In this work, considering that RNN and CNN have their own advantages in skeleton based action recognition, we construct a deep neural network, through merging the RNN with CNN.

To be specific, Bidirectional gated recurrent units (BIGRU) is used to achieve attention mechanism, and convolutional network for classification. An ensemble network including three subnets with the same structure is presented to learn diverse features for better accuracy. With the proposed method assessed on two classic benchmark datasets, namely, SBU Interaction Dataset [9] and NTU RGB + D Dataset [10], promising performance is achieved.

The main contributions of this paper are listed in the following two aspects:

1. A multi-feature representation method of interaction skeleton sequence is proposed for extracting various and complementary features. Specifically, three subnets fed with these features are fused into an ensemble network for recognition.
2. A framework combining RNN with CNN is designed for skeleton based interaction recognition, which can model the complex spatio-temporal variations in skeleton joints.

2. Related Work

In this section, the work related to the proposed method is briefly reviewed, including RNN based methods and CNN based methods for skeleton-based 3-D action recognition and interaction recognition.

RNN based methods: Some previous works have successfully applied RNN to skeleton based action recognition [11–13]. In [11] Du et al. divided the whole skeleton body into five parts according to the physical structure of human body, and fed them into five bidirectional LSTM jointly to make the decision of recognition. Zhu et al. [12] proposed an end-to-end fully connected deep LSTM network with a novel regularization scheme to learn the co-occurrence features of skeleton joints. In addition, they applied a new dropout algorithm to train the network. Liu et al. [13] proposed a spatio-temporal LSTM network which can model both temporal and spatial information. Based on LSTM, Song et al. [14] proposed a spatio-temporal attention model, which can automatically focus on the discriminative joints and pay different attention weights to each frame.

CNN based methods: Some previous works have employed CNN for skeleton based action recognition and achieved great success [15–20]. Ke et al. [19] represented the sequence as three clips for each channel of the 3D coordinates, which reflects the temporal information of the skeleton sequence and spatial relationship. Li et al. [21] proposed multiple views from skeleton sequences to learn the discriminative features including spatial domain feature and temporal domain feature, and multi-stream CNN fusion method was adopted to combine the recognition scores of all views. To exploit the spatio-temporal information from skeleton sequences, Kim and Reiter [22] used temporal convolutional neural networks (TCN) for skeleton based action recognition, which provided a way to explicitly learn readily interpretable spatio-temporal representations for 3D human action recognition.

The research of human action recognition brought about many surprising results with the development of RGB-D sensors [23]. However, few works talked about the human interaction recognition. Compared with single person's action recognition, two persons' interaction recognition is more complex and difficult [24,25]. Some early works [24] proposed a scheme of decomposing human interaction into single person's actions for recognition. Actually, lots of features in human interaction behavior which include both individual action information and mutual relations which can be utilized to obtain better recognition results.

3. Proposed Method

As shown in Figure 1, the proposed basic framework of the subnet is composed of two models: attention module and classification module. The input skeleton sequence consists of multiple frames, one column vector in the image matrix denotes one frame. Every frame consists of 3-dimensional joint coordinates. We separate these coordinates to x, y, z dimensions, which mean the R, G, B channels, respectively. For each channel, attention mechanism is used to learn the temporal weights of frames. After that, the three channels are concatenated to one tensor which is fed into classification module for classification. This section is organized as follows. Firstly, we introduce some different processes of

transforming skeleton sequences to color images with RGB three channels, and these images are used as different inputs of three subnetworks. Then the attention mechanism in action recognition is presented. Finally, an ensemble network with attention mechanism is constructed for interaction recognition.

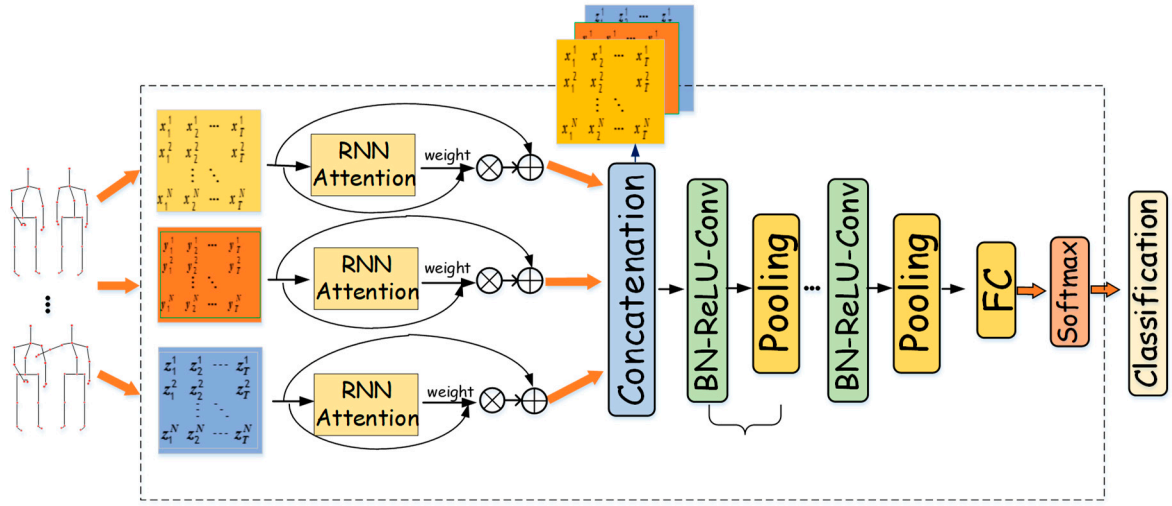


Figure 1. Basic framework of the subnet.

3.1. Multi-Feature Representation from Skeleton Sequences

Like the previous work in [26,27], the coordinates of the joint in one skeleton sequence can be arranged in a matrix, and three coordinates x, y, z of each joint represent the corresponding channels R, G, B of each color image. Then the coordinates of the j-th joint in each frame can be denoted as Equation (1)

$$P_j = \left(P_{xj}, P_{yj}, P_{zj}\right) \tag{1}$$

where $j \in \{1,\ 2,\ \cdots,\ m\}$, and m denotes the number of joints in each frame. For each frame, assume that there are two subjects and each subject has m joints. Then there are total $2 \times m$ joints for all subjects in each frame, let the j-th joint of the i-th performer map be P^i_j, and these joints at the t-th frame can be represented as Equation (2)

$$P_t = \left\{P^1_1, P^1_2,\ \cdots,\ p^1_m, P^2_1, P^2_2,\ \cdots,\ p^2_m\right\} \tag{2}$$

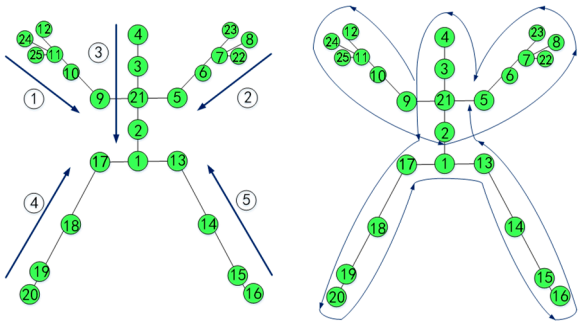
In order not to destroy the relation among these joints, the joints are numbered in a fixed order which determines the arrangements of all the joints. Considering that the human skeleton consists of such five parts as: one trunk, two arms, and two legs, we adopt two kinds of orders for skeleton arrangement: part-based order and traversal-based order [26]. Figure 2 shows two different orders on NTU RGB + D dataset [9] where one skeleton is composed of 25 joints. Then the whole interaction sequence can be represented as Equation (3)

$$P = [P_1,\ P_2,\ \cdots,\ P_T] \tag{3}$$

where T denotes the number of frames in an interaction sequence.

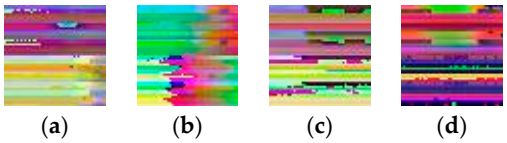
These coordinate elements can be regarded as RGB elements in images. In this way, we transform the original skeletal data to 3D tensors which can be sent to neural networks for training. We process the converted skeleton data to three different features for better performance.

Feature 1: we separated the two subjects and arrange them on one matrix as an image. For each subject, we adopted the part-based order. Figure 3 shows some feature images generated from sample actions.



(**a**) (**b**)

Figure 2. Different orders for skeleton arrangement. (**a**) part-based order (**b**) traversal-based order.



(**a**) (**b**) (**c**) (**d**)

(**e**) (**f**) (**g**) (**h**)

Figure 3. Sample feature images generated on NTU RGB+D dataset. (**a**) Kicking other person; (**b**) pushing other person; (**c**) patting on back of other person; (**d**) pointing finger at the other person; (**e**) hugging other person; (**f**) giving something to other person; (**g**) touching other person's pocket; (**h**) handshaking.

Feature 2: for different kinds of interaction, the relationship information between two subjects is distinct, which can be represented by the distances between the values at corresponding skeleton joints of two subjects in the same frame. Let $P_i^1(t)$ and $P_i^2(t)$ denote the i-th joint coordinates of the first and second player at the t-th frame. The Euclidean distance $D_i(t)$ between the corresponding joints of two subjects is denoted as Equation (4)

$$D_i(t) = \left\| P_i^2(t) - P_i^1(t) \right\| \text{ i} \in \{1, 2, \ldots, m\} \text{and } t \in \{1, 2, \ldots, T\} \tag{4}$$

where m denotes the number of joints and T refers to the total number of frames. In this feature mode, we adopt traversal-based order. Then an interaction instance D of all T frames can be represented as Equation (5)

$$D = \left\{ D_1, D_2, \cdots, D_T \right\} \tag{5}$$

Feature 3: we enhance the relationship information represented by the distances between the values at different skeleton joints in two subjects in the same frame. The enhanced cross joint distances $D_{ij}(t)$ between joint j and i of two performers at frame t can be represented as Equation (6)

$$D_{ij}(t) = \left\| P_j^2(t) - P_i^1(t) \right\| \text{ i}, j \in \{1, 2, \ldots, m\} \tag{6}$$

where i and j denote the joint number of the performers independently. For this feature, there are $m \times m$ joint distances for each frame. Then an interaction instance D' of T frames can then be represented as Equation (7)

$$D\prime = \left\{ D_1', D_2', \cdots D_T' \right\} \tag{7}$$

It can be seen all these features include single person's action feature as well as the relationship of human interaction.

3.2. Attention Mechanism

In terms of human attention mechanism, we design an attention mechanism in our ensemble network. Human usually focus on specific parts they are interested in for their visual subject and critical moments when behavior occurs. The skeleton data are a time sequence of multi-frame 3D joint coordinates forming an action. For different frames, it is of different levels of significance for recognition. For example, for the interaction punching and handshaking, the actions in most of the frames are similar so we should pay more attention on the key frames which carry more effective information. Inspired by the attention mechanism [14,28], we design an attention mechanism where each frame is assigned a different attention weight in order to emphasize key frames which contain important and discriminative information.

The learning of attention mechanism pursuits a specific attention based on BiGRU in memory cell to capture the temporal memory information across the input interaction sequence. As is shown in Figure 3, the output of BiGRU is determined by the forward GRU and backward GRU, so it can pay specific attention on the skeleton sequence by the context information. More specifically, the output of the attention module can be represented by Equation (8)

$$F(X) = X \circ F_A(X) \tag{8}$$

where X is a column vector in Equations (3), (5), and (7) for three features, which means one frame in the action, $F_A(X)$ is the weight of the frame vector X to enhance temporal information, and $\circ$ refers to element-wise multiplication. $F_A(X)$ can be computed as Equation (9)

$$F_A(X_t) = \sigma\left(\overrightarrow{GRU}(X_t) + \overleftarrow{GRU}(X_t)\right) \tag{9}$$

where $\sigma(\cdot)$ refers to sigmoid activation function, $\overrightarrow{GRU}(x_t)$ and $\overleftarrow{GRU}(x_t)$ denote the hidden variables of the forward GRU and backward GRU at t frame. The attention module can automatically learn the attention weight F_A of different frames from the output $F_A(X_t)$ in BiGRU. Among these frames, the larger the value of activation function, the more important this frame is for determining the category of interaction.

3.3. Ensemble Network

Our network consists of two modules: bidirectional gated recurrent units for attention module and convolutional neural networks for classification module. For the attention module, the number of units in BiGRU is set to be 128, and the recurrent dropout rate is set to be 0.5.

By utilizing the robustness of CNN to deformation, high-level feature representations can be extracted in the classification to better cope with spatio-temporal variations of skeleton joints. In principle, any CNN can be used in classification module, e.g., DenseNet and ResNet. In our method, we use the AlexNet [15] as our basic convolutional network, which is a very simple but effective network structure. Figure 4 shows the proposed convolutional module. We stack 3 Conv-ReLU-BN blocks. The convolutional strides and pooling strides are (2, 2). In convolutional layer, we use ReLU activation function. After the blocks, we add dropout layer and two FC layers, the number of the units for the last FC layer (i.e., the output layer) is the number of the action classes in each dataset. With different features as inputs of three subnetworks, we train these subnetworks both independently and globally. Cross-entropy is taken as the cost function, which can be described as Equation (10)

$$Loss = -\sum_{i=1}^{n} y_i \log(\hat{y}_i) \tag{10}$$

where y_i is the one hot vector of true label, $\hat{y}_i$ is the prediction vector, and n is the number of interaction classes.

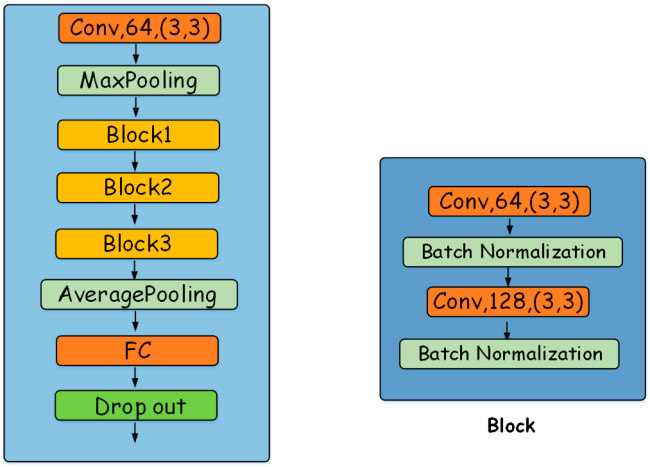


Figure 4. Convolutional module. "conv, 64, (3,3)" refers to one convolutional layer with 64 filters and the kernel size is 3×3.

The ensemble network framework is shown in Figure 5. We train the ensemble network end to end, all outputs of the three subnetworks are joined to determine the recognition result of interaction classes. We apply two fusion methods [29] in our ensemble network.

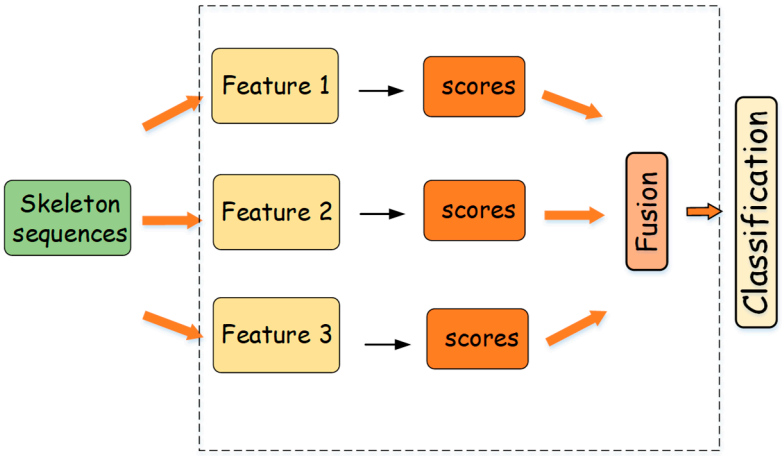


Figure 5. Ensemble network framework.

(i) Product fusion

Based on the product rule, the three subnetworks' output score vectors are element-wise multiplied. Based on the highest score, the predicted class can be expressed as Equation (11)

$$class = argmax\ (v_1 \circ v_2 \circ v_3) \tag{11}$$

where v is a score vector of three subnetworks' outputs, $\circ$ means element-wise multiplication, and $argmax(\cdot)$ refers to looking for the class index of the element with the highest score.

(ii) Sum fusion

In the similar way of the afore-mentioned product rule, the input to class label is assigned as Equation (12) through the sum rule

$$class = argmax(v_1 + v_2 + v_3) \tag{12}$$

where $+$ denotes element-wise addition.

4. Experimental Results

4.1. Dataset

In the following experiments, we assessed our proposed method on two public widely used datasets: SBU Interaction Dataset [9] and NTU RGB + D Dataset [10].

SBU-Kinect dataset. This dataset is a human interaction recognition dataset captured by Kinect and depended by two-person interaction. It contains 282 skeleton sequences and 6822 frames of eight classes. There are 15 joints for each skeleton. For fair comparison, we adopt five-fold cross validation protocols as suggested in [9].

NTU RGB + D Dataset. The NTU dataset is a high-quality action recognition dataset consisting of more than 56,000 action samples. It provides 3-dimensional coordinates of skeleton joints. There are total 60 classes of actions carried out by 40 subjects, where the ratio of interaction behaviors to all classes of action behaviors is 11/60. The large variations on viewpoint, intra class and sequence length determine its demandingness. In fairness, we follow the standard cross-subject and cross-view evaluation protocols in [10].

4.2. Implementation Details

For the original NTU RGB+D dataset, we transposed the original coordinate system to human-centric coordinate system. Different from [30], we always chose the first person's body center as the center of the coordinate system in order to better express the relative position between two subjects. Furthermore, coordinate transformation can eliminate the influence of different perspectives of actions. Figure 6 shows the proposed human-centric coordinate system. The formula of calculating transformation of coordinates is shown as (13)

$$\vec{H'} = \vec{H} \times R + C \tag{13}$$

where $\vec{H}$ and $\vec{H'}$ are the original coordinate and the converted coordinate, and C is the coordinate of the body center of the first person. R is the rotation matrix.

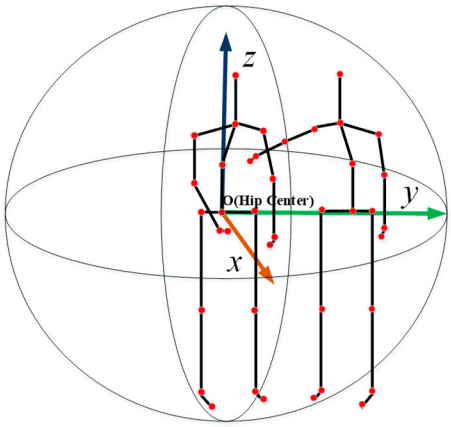


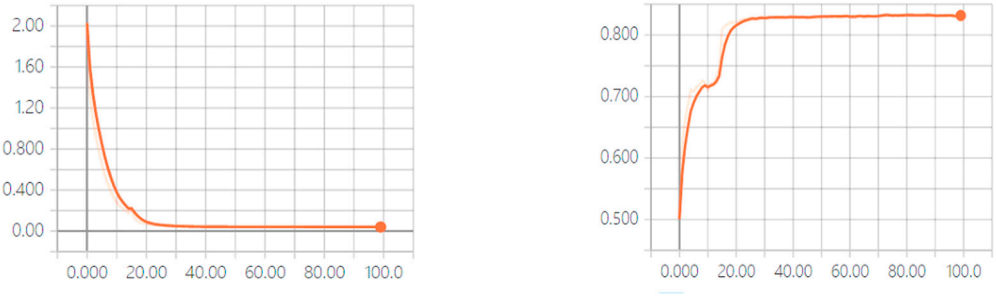
Figure 6. Human-centric coordinate system.

For the NTU RGB + D datasets, the matrices were obtained from all the frames of a skeleton sequence, since each person has $m = 25$ body joints and every interaction was considered to be acted by two subjects. When denoting single person's actions, we considered the second performer's joint coordinates were always zeros. Then in each frame there would be $N = 2 \times m$ joints, so the original image size was $3 \times N \times T$ for the feature 1 using the part-based arrangement, where T is the length of the sample. For feature 2, we adopted traversal-based arrangement and the size of the generated image was $3 \times N \times T$. For feature 3, we chose 16 rather than 25 key body joints for NTU RGB + D datasets to

decrease the amount of calculation and reduce model complexity. Then the size of image was $3 \times 256 \times T$. In order to meet the input requirements of the network, we fixed the length T of images by scaling the column number of the matrix from t to fixed t' through a bilinear interpolation scheme. Some early works [26] confirmed it was better to resize the images to a square size for recognition. Therefore, we resized the image to $3 \times 50 \times 50$, $3 \times 50 \times 50$ and $3 \times 256 \times 256$ for three features, respectively.

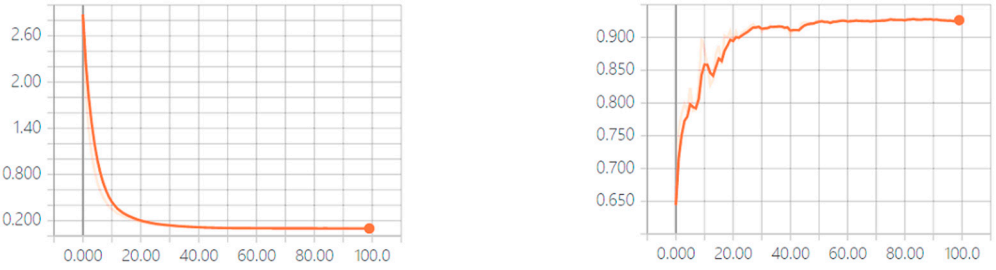
Compared with the above datasets, we used a similar method on SBU dataset, and resized the image to $3 \times 30 \times 30$, $3 \times 30 \times 30$ and $3 \times 225 \times 225$ for three features respectively since the SBU interaction dataset has less body joints and shorter interaction sequences.

For the training of the model, stochastic gradient descent algorithm with Nesterov acceleration with a momentum of 0.8 was adopted for optimization. The initial learning rate was set as 0.01, and decreased by a factor of 0.1 every 25 epochs. The batch size was 64 and the dropout rate was 0.3. After 100 epochs, the training process stopped. Figures 7–9 show the training loss and test accuracy curves of the best performance of our methods acquired by product fusion for NTU cross-subject, cross-view protocols and SBU dataset independently. As can be seen, the convergence speed was very fast.



(**a**) (**b**)

Figure 7. Training and test curve on NTU dataset (cross-subject): (**a**) training loss curve; (**b**) test accuracy curve.



(**a**) (**b**)

Figure 8. Training and test curve on NTU dataset (cross-view): (**a**) training loss curve; (**b**) test accuracy curve.

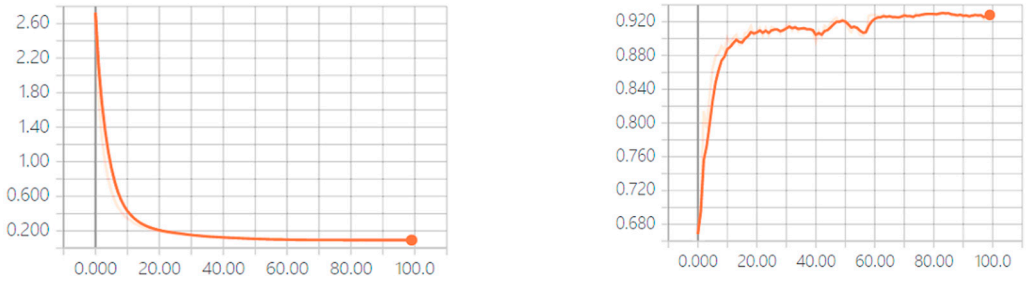
4.3. Results

Experimental results of the three subnets and ensemble-networks on two datasets have been listed in Table 1.

For NTU RGB+D dataset, it can be seen that all these subnetworks achieved good performances for both cross-subject and cross-view evaluation protocols based on our methods. On cross-view evaluation protocols, our method performed better due to the less variety among action performers.

Furthermore, the human-centric coordinate system could eliminate this influence of different perspectives of actions, which verifies the availability of coordinate transformation. The best performance was achieved by feature 3, because it carried more information. Furthermore, the score

fusion strategy improved the final accuracy by almost 3%, and product fusion method performed better than sum fusion, which exhibits the effectiveness of our approach.



(**a**) (**b**)

Figure 9. Training and test curve on SBU dataset: (**a**) training loss curve; (**b**) test accuracy curve.

Table 1. Recognition accuracy on NTU RGB+D and SBU datasets.

Feature	Dataset		
	NTU RGB+D		SBU
	Cross Subject	Cross View	
Feature 1	77.52%	86.73%	89.28%
Feature 2	77.95%	86.26%	89.94%
Feature 3	79.86%	87.33%	92.25%
Sum fusion	81.49%	90.12%	93.23%
Product fusion	82.53%	91.75%	93.58%

For SBU dataset, our networks also achieved relatively better performance with three subnetworks and ensemble network, the fusion strategy improved the accuracy from 92.25% to 93.58%, which proves the generalization ability of our method.

Table 2 shows the comparison result between our method and other methods on SBU dataset. Compared with other methods including hand-crafted feature-based methods and deep learning method, our approach achieved comparable performance except for [27,31]. Reference [27] generated different clips from skeleton sequences and proposed a multitask convolutional neural network to learn the generated clips and achieved 94.17% accuracy, which led to the increase of computation complexity and time consumption. In [31], Li et al. proposed an end-to-end convolutional co-occurrence feature learning framework hierarchical aggregation which could encode the spatial and temporal contextual information simultaneously, and achieved the state-of-the-art results. However, our proposed model had fewer layers and thus required fewer parameters than [31].

Table 2. Performance comparison of different methods on SBU dataset.

Method	Accuracy
Raw Skeleton [9]	49.70%
Joint Feature [32]	86.90%
CHARM [33]	83.90%
Hierarchical RNN [11]	80.35%
Deep LSTM [12]	86.03%
Deep LSTM+Co-occurrence [12]	90.41%
ST-LSTM [13]	88.60%
ST-LSTM+Trust Gate [13]	93.30%
RotClips+MTCNN [27]	94.17%
HCN [31]	98.60%
Proposed Method	93.58%

Table 3 lists the performance comparison of the proposed method with other state-of-the-art approaches for the NTU dataset; we can see our proposed model achieved excellent performances of 82.53% and 91.75%. Especially on cross-view evaluation protocols, our method performed better than others, which demonstrates the effectiveness of coordinate transformation system. On cross-sub evaluation protocols, our method also achieved good results, however, there were some gaps with the state-of-the-art method. One reason is that our method was mainly about human interaction recognition, the features of single person' actions got weakened due to the side effect of zero padding, which affected our recognition results.

Table 3. Performance comparison of different methods on NTU RGB + D dataset

Method	Cross Subject	Cross View
Hierarchical RNN [11]	59.10%	64.00%
Dynamic skeletons [34]	60.23%	65.22%
ST-LSTM+Trust Gate [13]	69.20%	77.70%
Two-stream RNNs [24]	71.30%	79.50%
STA-LSTM [30]	73.40%	81.20%
Res-TCN [22]	74.30%	83.10%
ST-GCN [35]	81.50%	88.30%
Multiview IJTM [21]	82.96%	90.12%
HCN [31]	86.50%	91.10%
Proposed Method	82.53%	91.75%

5. Conclusions

In this paper, we propose an ensemble network for skeleton based interaction recognition. In our model, diverse and complementary features are extracted from the original skeleton data as the inputs of three sub-networks. The three subnets are fused as one ensemble network. To learn different levels of significance of different frames adaptively, we design an attention mechanism based on BiGRU where each frame is assigned a different attention weight in order to emphasize key frames which contain important and discriminative information. Excellent results have been achieved on two widely used datasets and the results have shown that our proposed method is effective for feature extraction and recognition.

However, the proposed method was only evaluated in human action recognition and interaction recognition. In the future, we will focus on multiple-person related group activity.

Author Contributions: Conceptualization, methodology and writing original draft preparation, X.W.; writing—review, editing and supervision, H.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by a Sichuan University research grant.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Niebles, J.C.; Fei-Fei, L. A hierarchical model of shape and appearance for human action classification. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 19–21 June 2007; pp. 1–8.
2. Niebles, J.C.; Wang, H.; Fei-Fei, L. Unsupervised learning of human action categories using spatial-temporal words. *Int. J. Comput. Vis.* **2008**, *79*, 299–318. [CrossRef]
3. Rezazadegan, F.; Shirazi, S.; Upcrofit, B.; Milford, M. Action recognition: From static datasets to moving robots. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017; pp. 3185–3191.
4. Jalal, A.; Kim, Y.H.; Kim, Y.J.; Kamal, S.; Kim, D. Robust human activity recognition from depth video using spatiotemporal multi-fused features. *Pattern Recognit.* **2017**, *61*, 295–308. [CrossRef]

5. Han, J.G.; Shao, L.; Xu, D.; Shotton, J. Enhanced computer vision with Microsoft Kinect Sensor: A review. *IEEE Trans. Cybern.* **2013**, *43*, 1318–1334. [PubMed]
6. Shotton, J.; Fitzgibbon, A.; Cook, M.; Sharp, T.; Finocchio, M.; Moore, R.; Kipman, A.; Blake, A. Real-Time human pose recognition in parts from single depth images. In Proceedings of the IEEE Conference Computer Vision and Pattern Recognition, Providence, RI, USA, 20–25 June 2011; pp. 1297–1304.
7. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
8. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Gated feedback recurrent neural networks. In Proceedings of the 32nd International Conference on Machine Learning, Lile, France, 6–11 July 2015.
9. Yun, K.; Honorio, J.; Chattopadhyay, D.; Berg, T.L.; Samaras, D. Two-person interaction detection using body-pose features and multiple instance learning. In Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012; pp. 28–35.
10. Shahroudy, A.; Liu, J.; Ng, T.T.; Wang, G. NTU RGB+D: A large scale dataset for 3-D human activity analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1010–1019.
11. Du, Y.; Wang, W.; Wang, H. Hierarchical recurrent neural network for skeleton based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1110–1118.
12. Zhu, W.; Lan, C.; Xing, J.; Zeng, W.; Xie, X. Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; pp. 3697–3704.
13. Liu, J.; Shahroudy, A.; Xu, D.; Wang, G. Spatio-Temporal LSTM with trust gates for 3D human action recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; Volume 9907, pp. 816–833.
14. Song, S.; Lan, C.; Xing, J.; Zeng, W.; Liu, J. An end-to-end spatiotemporal attention model for human action recognition from skeleton data. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; pp. 4263–4270.
15. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
16. He, K.M.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
17. Li, C.; Hou, Y.; Wang, P.; Li, W. Joint distance maps based action recognition with convolutional neural network. *IEEE Signal Process. Lett.* **2017**, *24*, 624–628. [CrossRef]
18. Xu, Y.; Cheng, J.; Wang, L.; Xia, H.; Liu, F.; Tao, D. Ensemble one- dimensional convolution neural networks for skeleton-based action recognition. *IEEE Signal Process. Lett.* **2018**, *25*, 1044–1048. [CrossRef]
19. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 4570–4579.
20. Ke, Q.; An, S.; Bennamoun, M.; Sohel, F.; Boussaid, F. SkeletonNet: Mining deep part features for 3-d action recognition. *IEEE Signal Process. Lett.* **2017**, *24*, 731–735. [CrossRef]
21. Li, C.; Hou, Y.; Wang, P.; Li, W. Multiview-based 3-D action recognition using deep networks. *IEEE Trans. Hum. Mach. Syst.* **2019**, *49*, 95–104. [CrossRef]
22. Kim, T.S.; Reiter, A. InterpreTable 3D Human Action Analysis with Temporal Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 1623–1631.
23. Wang, P.; Li, W.; Ogunbona, P.; Wan, J.; Escalera, S. RGB-D-based human motion recognition with deep learning: A Survey. *Comput. Vis. Image Underst.* **2018**, *171*, 118–139. [CrossRef]
24. Ji, Y.; Cheng, H.; Zheng, Y.; Li, H. Learning contrastive feature distribution model for interaction recognition. *J. Vis. Commun. Image Represent.* **2015**, *33*, 340–349. [CrossRef]

25. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 3633–3642.
26. Cao, C.; Lan, C.; Zhang, Y.; Zeng, W.; Lu, H.; Zhang, Y. Skeleton- based action recognition with gated convolutional neural networks. *IEEE Trans. Circuits Syst. Video Technol.* **2018**. [CrossRef]
27. Ke, Q.; Bennamoun, M.; An, S.; Sohel, F.; Boussaid, F. Learning clip representations for skeleton-based 3d action recognition. *IEEE Trans. Image Process.* **2018**, *27*, 2842–2855. [CrossRef] [PubMed]
28. Pei, W.; Baltrusaitis, T.; Tax, D.M.; Morency, L.P. Temporal attention-gated model for robust sequence classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 820–829.
29. Jain, A.; Nandakumar, K.; Ross, A. Score normalization in multimodal biometric systems. *Pattern Recognit.* **2005**, *38*, 2270–2285. [CrossRef]
30. Liu, B.; Ju, Z.; Liu, H. A structured multi-feature representation for recognizing human action and interaction. *Neurocomputing* **2018**, *318*, 287–296. [CrossRef]
31. Li, C.; Zhong, Q.; Xie, D.; Pu, S. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; pp. 786–792.
32. Ji, Y.; Ye, G.; Cheng, H. Interactive body part contrast mining for human interaction recognition. In Proceedings of the 2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), Chengdu, China, 14–18 July 2014; pp. 1–6.
33. Li, W.; Wen, L.; Chuah, M.C.; Lyu, S. Blind human action recognition: A practical recognition system. In Proceedings of the IEEE International Conference on Computer Vision, Columbus, OH, USA, 24–27 June 2014; pp. 4444–4452.
34. Ohn-Bar, E.; Trivedi, M. Joint angles similarities and HOG2 for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Portland, OR, USA, 23–28 June 2013; pp. 465–470.
35. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.