doi:10.3772/j.issn.1006-6748.2023.04.007

Two stream skeleton behavior recognition algorithm based on Motif-GCN①

WU Jin(吴 进)②, WANG Lei, FENG Haoran, CHONG Gege
(School of Electronic Engineering, Xi'an University of Posts and Telecommunications, Xi'an 710121, P. R. China)

Abstract

Compared with RGB videos and images, human bone data is less vulnerable to external factors and has stronger robustness. Therefore, behavior recognition methods based on skeletons are widely studied. Because graph convolution network (GCN) can deal with the irregular topology data of human skeletons very well, more and more researchers apply GCN to human behavior recognition. Traditional graph convolution methods only consider the joints with physical connectivity or the same type when building the behavior recognition model based on human skeletons structure, which cannot capture higher-order information better. To solve this problem, Motif-GCN is used in this paper to extract spatial features. The relationship between the joints with natural connection in the human body is encoded by the first Motif-GCN, and the possible relationship between the unconnected joints in the human skeleton is encoded by the second Motif-GCN. In this way, the relationship between non-physical joints can be strengthened. Then a two stream framework combining joint and bone information is used to capture more action information. Finally, experiments are conducted on two subdatasets X-Sub and X-View of NTU-RGB + D, and the accuracy shown in Top-1 classification results is 89.5% and 95.4% respectively. The experimental results are 1.0% and 0.3% higher than those of the 2S-AGCN model respectively. The superiority of this method is also proved by the experimental results.

Key words: skeleton behavior recognition, Motif-GCN, two stream network

0 Introduction

Human behavior recognition is of great research significance in computer vision. In essence, it enables computers to recognize human behavior intelligently. Based on this characteristic, it is widely used in video surveillance[1], intelligent transportation[2], human-computer interaction, virtual reality simulation[3], intelligent security [4] and smart home, etc.

Compared with the methods based on RGB (red green blue) images or videos, the behavior recognition based on human skeletons data is less susceptible to the influence of external factors and has better robustness. Therefore, it has been widely studied. Traditional deep learning-based skeletons behavior recognition methods convert human skeletons data into vector sequences or two-dimensional grids, and then input them into convolutional neural network (CNN) or recurrent neural network (RNN) for prediction. However, the graph structure of human skeletons data itself is neglected in this method, and the dependence relationship between related joints cannot be represented fully. However, the graph convolution network (GCN) emerged in recent years can well deal with the data with irregular topology structure. In addition, the continuous maturation and development of equipment capturing human skeletons coordinates in recent years, such as Openpose, Optical Camera, Microsoft Kinect, Intel RealSense. As a result, human skeletons behavior recognition based on GCN has been widely studied. In 2018, Yan et al. [5] proposed the application of GCN to skeletal behavior recognition. They regarded human joint nodes as graph nodes, human joints naturally connected and the same joints in continuous frames as edges to construct spatio temporal graph convolutional network (ST-GCN). However, ST-GCN has limitations in the process of graph construction. (1) The graph in ST-

① Supported by the National Natural Science Foundation of China (No. 61834005, 61772417, 61802304) and the Shaanxi Province Key Research and Development Project (2021GY-280).
② To whom correspondence should be addressed. E-mail: wujin1026@126.com.
Received on Nov. 24, 2022

GCN only represents the physical structure of the human body, so it cannot better identify some human actions. For example, the action of putting on shoes, it is a very important human activity, but ST-GCN is difficult to capture the relationship between hands and feet. (2) ST-GCN only contains the first-order information of the human skeletons, namely the joint information, while the second-order information which also contains important action information of the skeletons is ignored. (3) The structure of GCN is hierarchical, and different layers contain semantic information of multiple layers. However, the graph topology in ST-GCN is fixed in all layers, which lacks the flexibility and ability to model the multi-level semantic information in all layers. Refs [6 – 14] and others on the basis of these problems have made the improvement. Aiming at problem (1) and problem (2), the use of Motif-GCN is proposed in this paper to extract spatial features of skeleton graph. The relationship between adjacent joint nodes that are naturally connected by the human body are mainly considered in the first Motif, and the relationship between joints that are not physically connected are considered in the second Motif.

The temporal features is extracted by the temporal convolutional network (TCN), the length and direction of the vector between two joints are regarded as the length and direction of the bone, and it's added to the GCN to predict the action label like the first-order information. Finally, the results of joint flow and bone flow are combined to obtain the final result.

1 Related work

1.1 Skeleton-based behavior recognition

The traditional behavior recognition based on skeletons mostly depends on the manual design features[15], however, the traditional way is based on the prior knowledge of researchers or data feature extraction, to some extent, the action characteristics of human behavior is often reflected, and it could not fully represent the overall state, and is susceptible to outside influences. With the rapid development of deep learning, behavior recognition based on deep learning has become a research hotspot. Common deep learning techniques include 3D convolutional neural network (3D-CNN) model [16], RNN model[17], two stream CNN model[18] and hybrid network model[19]. CNN-based methods convert skeleton data into pseudo-images based on hand-designed transformation rules, representing temporal dynamics and skeleton joints in the form of rows and columns, while RNN-based methods convert skeleton data into coordinate vector sequences. 3D-CNN model is to stack some frames into a cube shape and uses 3D convolution kernel for feature extraction. The idea of two stream convolution network is to input RGB information and optical flow field information of video frames into a CNN grid respectively, and then make prediction respectively. Finally, the final result is obtained by fusing the two prediction results. Hybrid network models include the combination of CNN and RNN, and the combination of CNN and long short term memory (LSTM). Spatial features are extracted by the former, while the temporal features are extracted by the latter. However, a large amount of data and parameters are required by the 3D-CNN, which is not conducive to the extraction of long-term features. The two stream method can only extract the temporal features of the before and after frames. The extracted temporal features are not very comprehensive, and the hybrid network model is difficult to combine them, with too many parameters and high resource consumption, which is difficult to deploy in reality. Since the skeleton data is the structure of a graph, it cannot be converted into vector sequences or two-dimensional grids to extract features well. Compared with CNN and RNN in recent years, GCN can handle this data structure well. In 2018, Yan et al. [5] proposed the application of GCN in behavior recognition based on human skeletons data, and a variety of improved methods have emerged since then. Based on Ref. [20], Motif-GCN is adopted to extract spatial features and TCN is adopted to extract temporal features. At the same time, the bone information is added to construct a two stream structure to further realize the purpose of strengthening the relationship between the joints of the body.

1.2 Graph convolutional network

Ref. [21] combined deep learning technology with graph data for the first time and graph neural network (GNN) is proposed, which made the deep learning technology be effectively used in the related scenes of graph data. Due to the success of CNN, the concept of CNN is generalized from grid data to graph data, and thus GCN is generated.

Construction on the graph of GCN method usually can be divided into the method based on spectral domain [21] and the method based on spatial domain[22]. GCN based on spectral domain is similar to the convolution theorem, the spatial domain of the signal through the Fourier transform to the spectral domain for multiplication operation, treatment after the transformation to the spatial domain, GCN method based on spectral domain is defined in the spatial domain of graph nodes signal through the graph spectral domain Fourier trans-

form into, and then, finally a method to transform back to the spatial domain. The method based on spatial domain is to directly process the graph signal. In the CNN convolution operation, each pixel is treated as a node, and the new features of the node are obtained by calculating the weighted average value of the neighbor nodes around the node. As for graph data, its structure is irregular, so it can not directly realize feature extraction by sliding convolution kernel like CNN. In GCN, the neighboring nodes of each node are usually sampled, and then these neighbor nodes are divided into different subsets to realize weight sharing and finally realize feature extraction.

2 Introduction to algorithm Principle

In the traditional method of behavior recognition based on human skeletons, in the graph structure constructed with human joints as nodes and bone as edges, only adjacent nodes of joints are generally considered when using GCN to extract spatial features. In this paper, Motif-GCN is used to extract spatial relationship between joints. The relationship between joints that are directly adjacent is encoded by the first Motif-GCN, and the relationship between joints that are disconnected is encoded by the second Motif-GCN, so as to strengthen the relationship between physically connected and non-physically connected joints and capture higher-order information. In addition, bone information is also introduced to construct the two stream structure of bone and joints.

2.1 Motif-GCN

Compared with CNN, the convolution kernel has translation invariance, and graph structured data does not have this property because of its unique structure. Therefore, the biggest difference between GCN and CNN is the definition of sampling function and weight function. In the previous methods, most of them chose to sample the first-order or second-order neighbor nodes of each node, as shown in Fig. 1, and divided them into root nodes, centripetal nodes, and centrifugal nodes according to their distance from the center of gravity and the distance from the root node to center of gravity.

As shown in Fig. 1, it is the traditional neighbor node sampling rules, where 0 represents the root node, 1 represents the centripetal node, 2 represents the centrifugal node, and the cross represents the center of gravity.

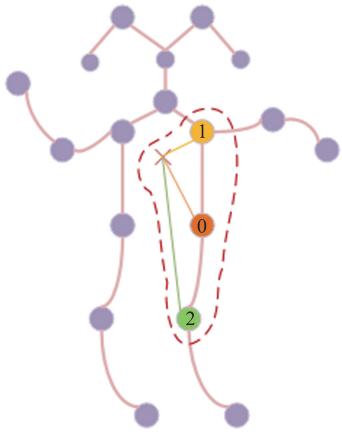

Fig. 1 Traditional neighbor node sampling rules

In the traditional graph convolution operation, the convolution operation at a node can be expressed as in Eq. (1)[5], where v_{ti} represents the central node, v_{tj} represents the neighbor node of v_{ti}, P represents the sampling function, W represents the weight function, and Z is shown in Eq. (2), Z represents the number of subsets divided by neighbor nodes, $l_{ti}(v_{tj})$ represents the label to which v_{tj} belongs in the molecular set label with v_{ti} as the center node, which is used to balance the contribution of different subsets.

$$f_{\text{out}}(v_{ti}) = \sum_{v_{tj} \in B(v_{ti})} \frac{1}{Z_{ti}(v_{tj})} f_{\text{in}}(P(v_{ti}, vtj))) \cdot W(v_{ti}, v_{tj}) \tag{1}$$

$$Z_{ti}(v_{tj}) = \{ v_{tk} \mid l_{ti}(v_{tk}) = l_{ti}(v_{tj}) \} \tag{2}$$

Although joints moved in groups when people perform movements, a single joint may appear in multiple parts of the body. In this sense, learnable mask M is added to each layer of spatio temporal graph convolution[5]. The mask will scale the contribution of a node's features to its neighbors based on the learned importance weight of each spatial graph edge in the edges naturally connected by the human body. Therefore, the graph convolution operation is finally represented as

$$f_{\text{out}} = \sum_{k}^{Kv} W_k (f_{\text{in}} \boldsymbol{A}_k) \odot M_k \tag{3}$$

In Eq. (3), f_{in} represents the input, $\boldsymbol{A}_k$ represents the adjacency matrix, M_k represents the learnable mask, and K_v represents the kernel size of the spatial dimension, $\odot$ denotes the dot product.

In 2019, Wen et al.[20] proposed Motif-GCN. Motif refers to the connection pattern between different node types, and a double Motif structure was constructed in Motif-GCN. The first Motif-GCN is used to encode the relationship between joints directly connected in the skeleton structure of human body, the relation-

ship between joints without connection in human skeleton structure is encoded by the second Motif-GCN. In the process of human movement, joints without connection often contain important action information. For example, in the movement of the right hand touching the left foot, although there is no connection between the hand and feet, the relationship between them is useful for identifying the movement of foot touching. Therefore, the addition of this Motif can achieve the capture of higher-order information.

The problem that the traditional graph convolution method for skeleton structure modeling only considers the physical connection neighbors of each joint and joints of the same type is solved by Motif-GCN, which cannot capture higher-order information. Different from Ref. [20], Motif-GCN is used for reference and Motif-GCN is choosed to extract spatial features while continuing to use TCN to extract temporal features. In addition, bone information was added in this paper to form a two stream structure with joint information. Then, the two were input into the 9-layers model composed of Motif-GCN and TCN respectively, and finally, Softmax classification function was used to obtain the classification results, and then a fusion of the two results was carried out to obtain the final classification verification results. The experimental results shown by the final combined results were significantly improved and enhanced compared with the previous classical models.

Two Motifs are used to simulate the physical connection and non physical connection of the human skeletons. In the first Motif, only the nodes with direct adjacent relationships are considered. The neighbor of each joint has three characters, the joint itself, the parent nodes of joint, and the child nodes of joint . It is shown in Fig. 2.

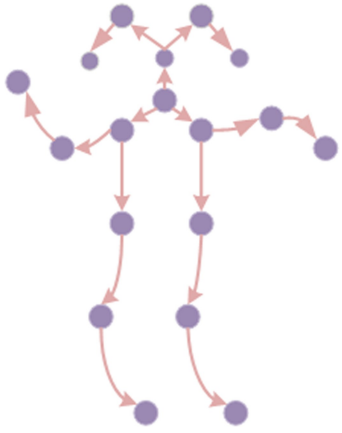

Fig. 2 Graph structure in the first Motif

As shown in Fig. 2, the graph structure in the first Motif is shown, where each circle represents a joint point, the line with arrow represents the natural physical connection of the human body, and the arrow represents from the parent node to the child node.

In the second Motif, the joints naturally connected by the human body are not considered, but the joints without physical connectivity are mainly considered. The weighted adjacency matrix between the disconnected joints is defined by allocating a large weight to the joints with short distance. In the weighted adjacency matrix, the relationship between node i and node j can be expressed as $\alpha_{i,j} = \max \boldsymbol{e} - \boldsymbol{e}(i,j)$, where $\boldsymbol{e}$ is a matrix representing the average euclidean distance between pairs of nodes in the sequence. The relationship between the joint that is not connected to the neck joint and the neck joint is shown in Fig. 3. The calculation formula of Euclidean distance between node 1 $(x_1, x_2, \cdots, x_n)$ and node 2 $(y_1, y_2, \cdots, y_n)$ in n-dimensional spaces is shown in Eq. (4).

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{4}$$

As shown in Fig 3, It is the structure between nodes in the second Motif. The two joints without physical connection is connected by the dotted line.

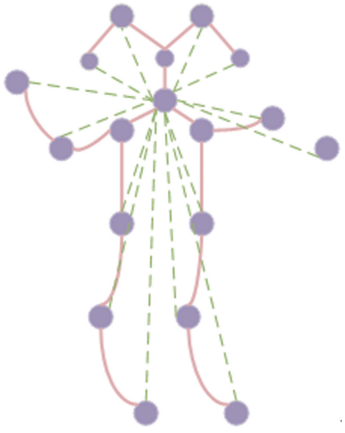

Fig. 3 Graph structure in the second Motif

Finally, Motif-GCN can be expressed as

$$Z_t^m = \sum_{k=1}^{K_m} (D_K{}^M)^{-1} \boldsymbol{A}_K{}^M X_t \boldsymbol{W}_K{}^M \tag{5}$$

In Eq. (5), $X_t \in R^{N \times D}$ represents the input. There are N nodes and $\boldsymbol{D}$ coordinates in frame t, and K_M represents the dependency between different semantics. Because the neighbor of each joint in the first Motif has three semantic roles, $K_{M1} = 3$, $K_{M2} = 1$; in the second

Motif, $\boldsymbol{A}_K^M$ represents the adjacency matrix corresponding to each Motif, and $\boldsymbol{D}$ is an angle matrix, where $\boldsymbol{W}_k^M$ represents the weight matrix corresponding to the node type k in each Motif, Z_t^M is output.

2.2 Overall network structure

The process of behavior recognition using Motif-GCN is shown in Fig 4. After bone sequence input, spatial features are extracted by Motif-GCN and temporal features are extracted by TCN. Finally, the final classification results are obtained through Softmax classifier. The specific algorithm process is shown in Fig. 4.

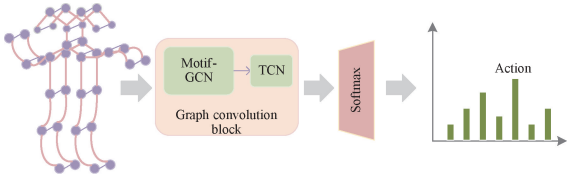

Fig. 4 The process of behavior recognition by Motif-GCN

As shown in Fig. 5, it is the overall flow chart of the Motif-GCN algorithm. Firstly, the skeleton sequence is input to obtain joint information and bone information, and then the two are fed into the Motif-GCN and TCN structures respectively. The Motif-GCN and TCN are followed by a batch normalization (BN) layer and ReLU layer. A Dropout layer is also added between the Motif-GCN and TCN. There are 9 layers in this structure, and the numbers below each layer represent the input channel number, output channel number and step size information of that layer. Then, the results are sent to the Softmax classifier to get the respective classification, Finally, the final classification results are obtained by result fusion.

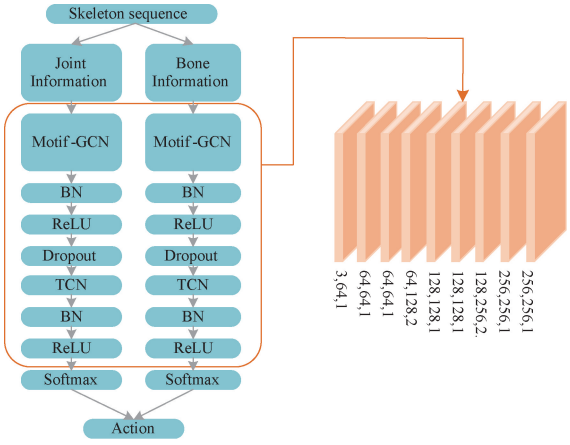

Fig. 5 Overall flow chart of two stream Motif-GCN algorithm

3 Experimental results and analysis

3.1 Dataset introduction

NTU-RGB + D[23] dataset consists of 56 880 action samples, including RGB video, 3D skeleton data, depth map sequence and infrared video for each sample. There are 60 categories, which are mainly divided into 3 groups, 40 daily activities, such as drinking, eating, reading; 9 health-related movements, such as sneezing, rocking and falling, and 11 reciprocal movements, such as answering, kicking and hugging. This data set was captured simultaneously by three Microsoft Kinect V. 2 cameras, with a resolution of 1920 × 1080 for RGB video, 512 × 424 for both depth map and infrared video, and 3D skeleton data containing the 3D positions of 25 major body joints per frame. NTU-RGB + D dataset adopts two different partitioning criteria when dividing training set and test set: Cross-Subject (X-Sub) and Cross-View (X-View).

Cross-Subject (X-Sub): in the cross-subject, a total of 40 subjects are used to collect data, and the age of these subjects is between 10—35 years old. This data set divides these 40 subjects into training group and testing group, with 20 subjects in each group, and the training set and testing set contain 40 320 and 16 560 samples respectively.

Cross-View (X-View): in the cross-view, three different horizontal views of − 45 ° and + 45 ° were captured from the same action using three cameras at the same height. Each subject was required to make each action twice, once facing the left camera and once facing the right camera. Two front views, one left view, one right view, a left 45 ° degree view and a right 45 ° view can be captured by this way.

Kinetics: data set contains about 300 000 video clips retrieved from YouTube. These videos cover up to 400 human action courses, from daily activities, sports scenes to complex interactive actions. In Kinetics, each clip lasts about 10. 240 000 videos are used to train and 20 000 videos are used for verification. Train the comparison model on the training set and report the accuracy on the verification set.

3.2 Experimental platform

As shown in Table 1, this experiment is carried out in Linux system, based on CUDA platform and combined with PyTorch deep learning framework. PyTorch version is 1. 1. 0, and GPU version is GTX 1080 Ti. The memory is Kingston HyperX Savage DDR4 and the hard drive is Seagate in 1 TB size.

3.3 Analysis of experimental results

The experiment was mainly conducted on the large data sets NTU-RGB + D and Kinetics. The accuracy and loss rate curves of the verification set on the NTU-RGB + D data set are shown in Figs. 6 – 13.

Table 1 Experimental platform

Name	Type(Version)	Instructions
OS	Linux	Ubuntu 18.10
CUDA	CUDA10.1	The underlying software platform for GPU acceleration
PyTorch	1.1.0	Dynamic graph characteristic
GPU	GTX 1080Ti	Memory 11 GB
Memory	Kingston HyperX Savage DDR4	Frequency 2400 MHz Memory 8GB
Hard disk	Seagate	1 TB

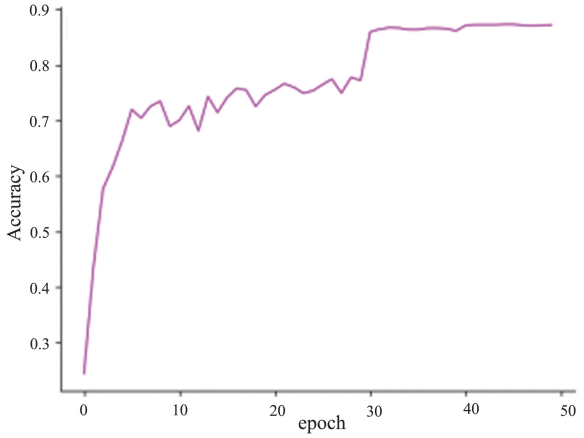

Fig. 6 X-sub joint accuracy

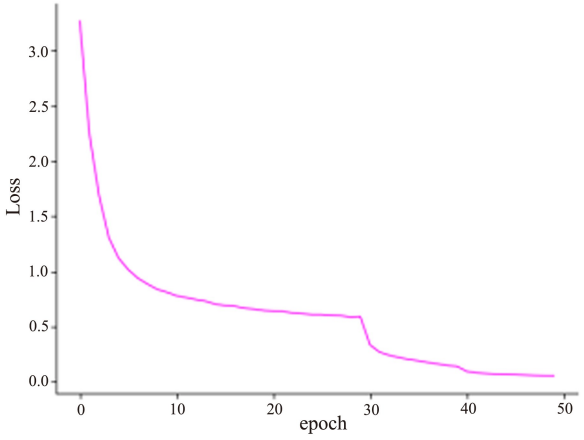

Fig. 7 X-sub joint loss

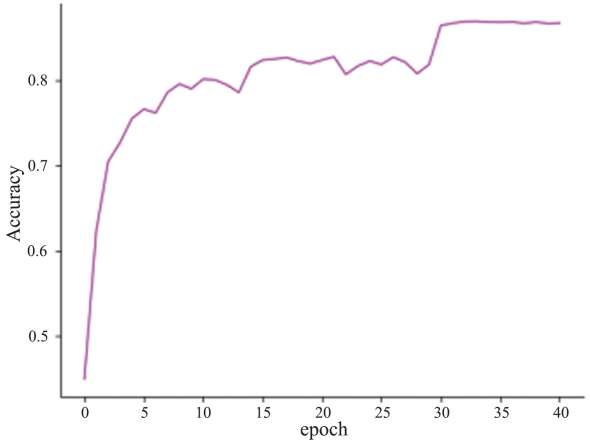

Fig. 8 X-sub bone accuracy

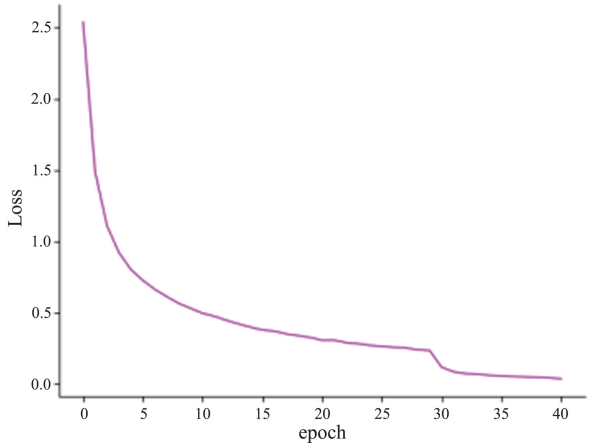

Fig. 9 X-sub bone loss

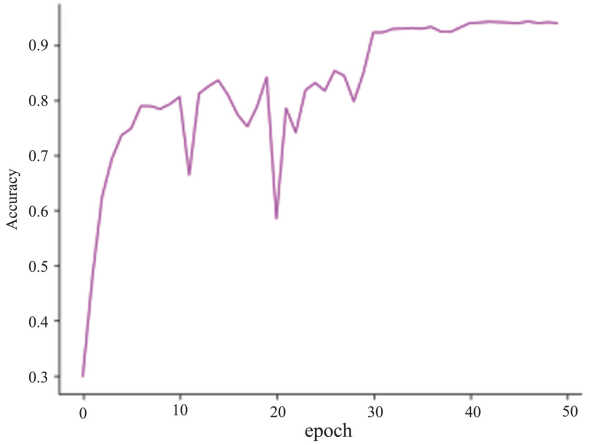

Fig. 10 X-view joint accuracy

As shown in Figs 6 – 13, in the X-Sub dataset, the accuracy and loss rate of joint flow and bone flow tend to be stable after epoch 30. The final accuracy rate of joint flow is about 0.873, the loss rate is about 0.489, the accuracy rate of bone flow is about 0.869, and the loss rate is about 0.508. Similarly, in the X-View data-

set, the accuracy and loss rate of joint flow and bone flow tend to be stable after epoch 30. The final accuracy rate of joint flow is about 0.942, the loss rate is about 0.197, the accuracy rate of bone flow is about 0.938, and the loss rate is about 0.203.

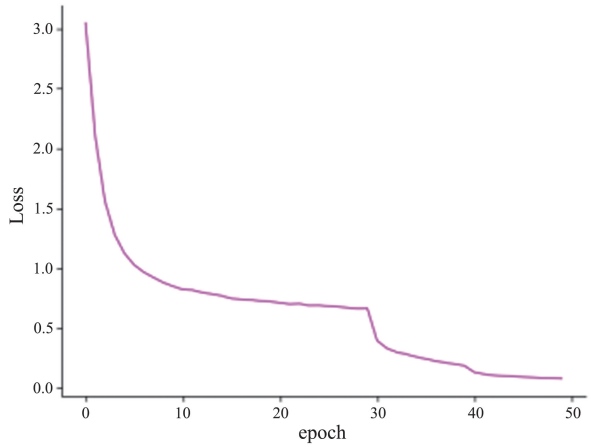

Fig. 11 X-view joint los

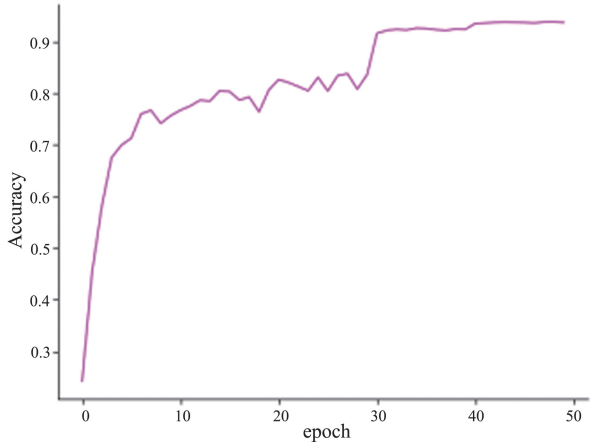

Fig. 12 X-view bone accuracy

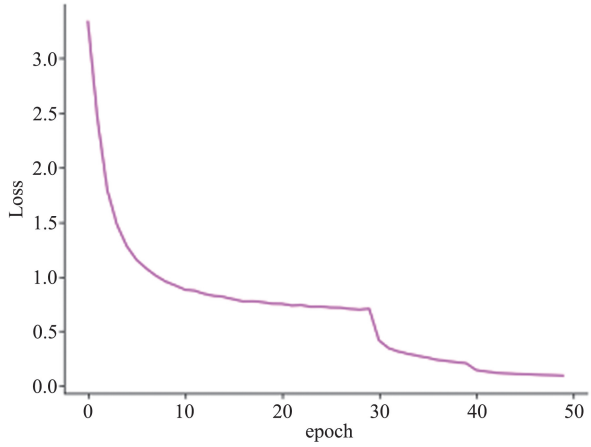

Fig. 13 X-view bone loss

The recognition accuracy on the validation set of NTU-RGB + D and Kinetics dataset are shown in Table 2 and Table 3. The comparison between the fused results and the results of other models is shown in Table 4.

Table 2 Experimental results under the NTU dataset

Dataset	Joint		Bone		Combined	
	Top1	Top5	Top1	Top5	Top1	Top5
X-Sub	87.3%	97.5%	86.9%	97.8%	89.5%	98.1%
X-View	93.7%	99.3%	93.8%	99.4%	95.4%	99.5%

Table 3 Experimental results under the Kinetics dataset

Dataset	Joint		Bone		Combined	
	Top1	Top5	Top1	Top5	Top1	Top5
Kinetics	34.3%	56.7%	34.4%	56.5%	36.7%	58.9%

Table 4 Comparison of validation accuracy between the proposed method and other methods on NTU-RGB + D and Kinetics dataset

Method	X-Sub	X-View	Kinetics	
	Top-1	Top-1	Top-1	Top-5
ST-GCN[5]	81.5%	88.3%	30.7%	52.8%
AS-GCN[24]	86.8%	94.2%	34.8%	56.5%
2S-AGCN[6]	88.5%	95.1%	36.1%	58.7%
GCN-NAS[8]	89.4%	95.7%	37.1%	60.1%
Proposed method	89.5%	95.4%	36.7%	58.9%

For the NTU-RGB + D dataset, each sample of the dataset has a maximum of 2 people. If there are less than 2 people in the sample, then 0 is used to fill. The maximum number of frames in each sample can be 300; if the number of frames is less than 300, repeat sampling until it reaches 300, the experiment is carried out on the PyTorch platform with stochastic gradient descent (SGD) algorithm. The batch size is 16, the weight attenuation is 0.001, and the learning rate is 0.1. The training will end at the 50th epoch. The experimental results in the two sub datasets of NTU-RGB + D of the method proposed in this paper is shown in Table 2. It can be seen from the Table 2 that the Top-1 accuracy of the joint results and bone results under X-Sub dataset is 87.3% and 86.9% respectively; the Top-1 accuracy of the joint results and bone results under X-View dataset is 93.7% and 93.8% respectively; and the final Top-1 accuracy of combined results of X-Sub and X-View is 89.5%, 95.4% respectively.

For Kinetics dataset, the experimental setup is the same as Ref. [6]. SGD algorithm is used in the experiment on PyTorch platform, with batch size of 8, weight attenuation of 0.001, and learning rate of 0.1. The training ends on the 65th epoch. The experimental results of the Kinetics dataset of the meth-

od proposed in this paper is shown in Table 3. It can be seen from the Table 3 that the Top-1 accuracy of the joint results and bone results under Kinetics dataset is 34.3% and 34.4% respectively; and the final Top-1 accuracy of combined results of Kinetics is 36.7%.

As shown in Table 4, compared with 2S-AGCN, the result of proposed method is improved by 1.0%, 0.3% respectively on X-Sub and X-View which are two subdatasets under NTU-RGB + D dataset, and is improved by 0.6% and 0.2% respectively on Kinetics dataset. The effectiveness of the method is also proved by this. However, compared with graph neural network neural architecture search (GCN-NAS), the accuracy of X-Sub is higher than that of GCN-NAS, while the accuracy of X-View and Kinetics is still not enough.

4 Conclusion

Aiming at the traditional behavior recognition method based on graph convolution, this paper only considers the problem of physical connection or the same type of joints when building the model, Motif-GCN is used to extract the spatial information of human skeleton points, The first Motif is used to encode the edges with natural connection relationship in the human body; the other Motif is used to encode the relationship between joints without connectivity in the human skeleton, and add joint and bone information at the same time. a two-stream structure is constructed, and experiments are carried out on the large dataset NTU-RGB + D. Finally, the accuracy rates on the two sub datasets X-Sub and X-View are 89.5% and 95.4% respectively, and the experimental results are 1.0% and 0.3% higher than those of the 2S-AGCN model. The method proposed in this paper, by adding the relationship between the joints of non physical connections and by building a two stream structure to add more action information, so as to strengthen the connection between the physical connection and non physical connection, joints in the human skeleton structure and captures the higher-order information. The effectiveness of this method is proved by the improvement of the experimental results compared with the 2S-AGCN model, but compared with some other recent methods, such as GCN-NAS, the experimental results need to be further improved.

References

[1] PATIL S, PRABHUSHETTY K S. An efficient motion based group level activity recognition for intelligent video surveillance[C]//2021 Asian Conference on Innovation in Technology. Pune: ASIANCON, 2021: 27-29.

[2] ALLADI T, KOHLI V, CHAMOLA V, et al. A deep learning based misbehavior classification scheme for intrusion detection in cooperative intelligent transportation systems[J]. Digital Communications and Networks, 2022, doi: https://doi.org/10.1016/j.dcan.2022.06.018.

[3] MARUYAMA R, TAKAHASHI S, HAGIWARA T. A virtual reality driving simulator with gaze tracking for analyzing driver's behavior[C]//2022 IEEE 4th Global Conference on Life Sciences and Technologies. Osaka: IEEE, 2022: 144-145.

[4] LIU Y C, ZHANG S N, LI Z Y, et al. Abnormal behavior recognition based on key points of human skeleton[J]. IFAC-Papers OnLine, 2020, 53(5): 441-445.

[5] YAN S J, XIONG Y J, LIN D H. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]// The 22nd AAAI Conference on Artificial Intelligence. New Orleans: AAAI Press, 2018: 1-10.

[6] SHI L, ZHANG Y F, CHENG J, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 12026-12035.

[7] ZHANG P F, LAN C L, ZENG W J, et al. Semantics-guided neural networks for efficient skeleton-based human action recognition[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 1112-1121.

[8] PENG W, HONG X P, CHEN H Y, et al. Learning graph convolutional network for skeleton-based human action recognition by neural searching[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Seattle: AAAI Press, 2020: 2669-2676.

[9] SONG Y F, ZHANG Z, SHAN C F, et al. Stronger, faster and more explainable: a graph convolutional baseline for skeleton-based action recognition[C]// Proceedings of the 28th ACM International Conference on Multimedia. Seattle: Association for Computing Machinery, 2020: 1625-1633.

[10] LI M S, CHEN S H, ZHAO Y H, et al. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 214-223.

[11] SONG Y F, ZHANG Z, WANG L. Richly activated graph convolutional network for action recognition with incomplete skeletons[C]//2019 IEEE International Conference on Image Processing (ICIP). Taibei: IEEE, 2019: 1-5.

[12] QIU H L, HOU B, REN B, et al. Spatio-temporal tuples transformer for skeleton-based action recognition[EB/OL]. (2022-01-08)[2022-11-04]. https://arxiv.org/pdf/2201.02849v1.pdf.

[13] CHENG K, ZHANG Y F, He X Y, et al. Skeleton-based action recognition with shift graph convolutional network[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 183-192.

[14] BAI R W, LI M, MENG B, et al. Hierarchical graph convolutional skeleton transformer for action recognition[C]//

2022 IEEE International Conference on Multimedia and Expo (ICME). Taibei: IEEE, 2022: 1-6.
[15] HORI R, HACHIUMA R, ISOGAWA M, et al. Silhouette-based 3D human pose estimation using a single wrist-mounted 360 ° camera [J]. IEEE Access, 2022, 10: 54957-54968.
[16] JIANG G H, JIANG X Y, FANG Z J, et al. An efficient attention module for 3d convolutional neural networks in action recognition[J]. Applied Intelligence, 2021, 51(10): 7043-7057.
[17] LIAO S J, LYONS T, YANG W X, et al. Logsig-RNN: a novel network for robust and efficient skeleton-based action recognition[EB/OL]. (2021-11-01) [2022-11-04]. https://arxiv. org/pdf/2110. 13008v2. pdf.
[18] CHEN G Z, YAO L, XU J T, et al. Two-stream adaptive weight convolutional neural network based on spatial attention for human action recognition [C] // International Conference on Intelligent Robotics and Applications. Harbin: Springer, 2022: 319-330.
[19] ZHU S M, GUENDEL R G, YAROVOY A, et al. Continuous human activity recognition with distributed radar sensor networks and CNN-RNN architectures [J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-15.
[20] WEN Y H, GAO L, FU H B, et al. Graph CNNs with motif and variable temporal block for skeleton-based action recognition [C] // Proceedings of the AAAI Conference on Artificial Intelligence. Hawaii: AAAI Press, 2019: 8989-8996.
[21] BRUNA J, ZAREMBA W, SZLAM A, et al. Spectral networks and deep locally connected networks on graphs[C] // The 2nd International Conference on Iearning Representations, Banff: ICLR, 2014: 1-14.
[22] SPUREK P, DANEL T, TABOR J, et al. Geometric graph convolutional neural networks [J]. Artificial Neural Networks, 2019, 26(3): 412-423.
[23] SHAHROUDY A, LIU J, NG T T, et al. NTU RGB + D: a large scale dataset for 3d human activity analysis [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 1010-1019.
[24] LI M S, CHEN S H, CHEN X, et al. Actional-structural graph convolutional networks for skeleton-based action recognition[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 3595-3603.

WU Jin, born in 1975. She received her B. S degree from Xi'an Jiaotong University in 1998, and she also received her M. S. degree from Xi'an Jiaotong University in 2001. Her research focuses on key techniques for signal and information processing.