Human action recognition model incorporating multiscale temporal convolutional network and spatiotemporal excitation network

Yincheng Qi[a,b,*] **Baoli Wang**[a] **Boqiang Shi**[a] **and Ke Zhang**[a,b]

[a]North China Electric Power University, School of Electrical and Electronic Engineering, Baoding, China

[b]North China Electric Power University, Hebei Key Laboratory of Power Internet of Things Technology, Baoding, China

Abstract. Human action recognition is a research hotspot in the field of computer vision. Focusing on the problem of similar action recognition, we propose an improved two-stream adaptive graph convolutional network for skeleton-based action recognition, which incorporating a multiscale temporal convolutional network and a spatiotemporal excitation network. Using the multiscale temporal convolutional network, the temporal information can be effectively extracted by dilated convolution at different scales so as to broaden the width of the temporal network and extract more temporal features with slight difference between categories at the same time. By utilizing the spatiotemporal excitation network, the input features can be obtained through channel pooling to form single-channel features for two-dimensional convolution, by which important spatiotemporal information can be excited and the role of local nodes in similar actions can be effectively enhanced. Extensive tests and ablation studies on the three large-scale datasets, NTU-RGB+D60, NTU-RGB+D120, and Kinetics-Skeleton, were conducted. Our model outperforms the baseline by 7.2% and the state-of-the-art model by 4% in the similar action recognition on NTU-RGB+D60 dataset on average, which demonstrates the superiority of our model. [DOI: 10.1117/1.JEI.32.3.033003]

Keywords: skeleton data; human action recognition; multiscale temporal convolutional network; similar action recognition; spatiotemporal excitation network.

Paper 221261G received Nov. 8, 2022; revised manuscript received Mar. 13, 2023; accepted for publication Apr. 17, 2023; published online May 4, 2023.

1 Introduction

With the rapid development of artificial intelligence, human action recognition has become an important research topic in computer vision, which is widely used in video surveillance, human–computer interaction, short video, virtual reality, and other fields.[1] A skeleton is a topological representation of the joints and bones of a human body, containing the connections between adjacent joints. It is robust to changes in human scale, visual angle, and movement speed, giving it an inherent advantage for action recognition based on human skeleton data. Accurate human skeleton data can be easily obtained in the context of the increasing maturity of depth sensors and human posture estimation technology, which makes it widely applied in practical projects.[2,3]

As deep learning technology progresses, data-driven approaches have become mainstream methods. The most widely used models are recurrent neural network (RNN) and convolutional neural networks (CNN).[4,5] Most of the RNN-based approaches are designed to solve the gradient disappearance and explosion problems and increase its spatial modeling capability.[6–9] CNN-based methods usually model the skeleton data as pseudoimages first from the perspective of data representation.[10–12] Later, graph convolutional network (GCN) generalized the convolution from images to graphs and was widely used in action recognition tasks based on skeleton data.[13] Many recent GCN-based methods have attempted to focus on the effective use of skeleton data.[14–17]

*Address all correspondence to Yincheng Qi, qiych@ncepu.edu.cn

To sum up, outstanding results have been achieved in terms of human action recognition in recent years, but the existing methods are less accurate for the recognition of similar actions. For example, they are not able to pay sufficient attention to the spatiotemporal information of similar actions, such as reading and writing, playing with phone versus type on a keyboard, etc. In practical, it is critical to be able to accurately recognize the actions with slight differences of various parts of the human body. Some actions in daily life are longer in duration and therefore more sensitive to temporal information, so it is also a key task to improve the recognition effect on long duration actions.

Inspired by this, we attempt to address the limitations of the above approaches in two ways. First, we use a two-stream adaptive graph convolutional network (2s-AGCN) as the basic framework to introduce the multiscale temporal convolutional network, aiming to achieve the extraction of the information about the difference of temporal features of nodes in three-dimensional (3D) skeleton data and broaden the overall width of the network at the same time.[14] Due to the different perceptual fields of different branches, the extraction of temporal features at different scales is achieved by setting two convolutional kernels with different expansion rates. Meanwhile, the depth of the training network is increased by the residual network. Second, a spatiotemporal excitation network for spatiotemporal information characterization in the video human action recognition task is used for Ref. 3. The spatiotemporal excitation network is utilized to focus on the spatiotemporal information of key local nodes. By utilizing the spatiotemporal excitation network, the important spatiotemporal information in the input features of each channel can be activated, and the role of key nodes is effectively enhanced in similar action recognition. Finally, we fuse the multiscale features with local features, and a new human action recognition model incorporating spatiotemporal excitation network and multiscale temporal convolutional network, spatiotemporal excitation multiscale graph convolutional network (STMGCN), is obtained. Compared with the baseline model, we replace the temporal convolutional network of 2s-AGCN with the multiscale temporal convolutional network and the residual connection with the spatiotemporal excitation network. To verify the effectiveness of the proposed network, extensive tests on three 3D skeleton datasets, NTU-RGB+D60, NTU-RGB +D120, and Kinetics-Skeleton, were conducted.[18–20] The obtained results demonstrate that the proposed method can not only effectively improve the accuracy of similar action recognition but also reduce the number of model parameters. It achieves a fine balance between accuracy and resource consumption in human action recognition.

The main contributions of this paper are summarized as follows. First, we used a multiscale temporal convolutional network to enhance the extraction ability of the temporal feature difference between categories, which not only improves the accuracy of similar action recognition but also reduces the number of model parameters. Second, we designed a spatiotemporal excitation network to make the spatiotemporal feature information of key local nodes in similar actions more prominent, which effectively improves the accuracy of action recognition. The proposed method outperforms most models in terms of recognition.

The rest of this paper is organized as follows. In Sec. 2, the related work is reviewed. In Sec. 3, the proposed method for human action recognition is introduced. The test results are given and discussed in Sec. 4. Finally, the conclusions are drawn in Sec. 5.

2 Related Work

Traditional human skeleton action recognition methods mainly use manually marked features to represent skeleton sequences and classify them according to the joint trajectories. However, the blindness and discrepancies are easily caused by manually designed features. The deep learning-based approaches automatically learn the data features and therefore retain more valuable information. Among them, the methods using RNN, CNN, and GCN are widely employed in human action recognition. RNN is able to process the sequence data efficiently and its improved versions, long short-term memory network (LSTM), and gated recurrent unit, etc., are more effective.[21,22] RNN-based action recognition methods usually improve the recognition accuracy by introducing linear correction units, such as independently recurrent neural network (IndRNN),[6] increasing the context dependence of spatial domain such as global context-aware

attention LSTM,[7] integrating short-term, medium-term, and long-term features in Ref. 8 named TS-LSTM and adding attention blocks.[9] Compared with RNNs, CNNs are able to learn the high-level semantic information more effectively and have advanced information extraction ability. Most of the CNN-based action recognition algorithms start from modeling the skeleton data as pseudoimage to satisfy the need of CNNs' input.[10–12] Compared with RNN, CNN is more suitable for processing skeleton data containing spatiotemporal information, which makes it more popular.

However, the dependencies between human joints cannot be sufficiently represented only using human skeleton data in non-Euclidean space as vector sequences or pseudoimages. GCN, as an extension of CNN, is an effective method to extract features from irregular graph data. Yan et al.[23] proposed a spatiotemporal graph convolutional network (ST-GCN) using human joints as nodes, natural connections between human joints in space as spatial edges, and connections of corresponding nodes of adjacent frames in time sequence as temporal edges to extract information in spatial and temporal dimensions. By this method, better recognition effect was obtained. In much of later study, ST-GCN was used as the base network and was improved. The most common research methods are the ones that utilize the skeleton data effectively. For example, Ref. 14 proposed the 2s-AGCN where the dual-stream network contains the first-order and the second-order information. Reference 15 proposed part-based graph convolutional network (PB-GCN), and the skeleton graph is divided into multiple parts based on the relationships between nodes. Reference 16 introduced a dynamic links module block named actional–structural graph convolution network (AS-GCN). In Ref. 17, the high-level semantics of joints are introduced into the network, and the method names semantics guided neural network (SGN). To expand the network width, some researchers used multiscale networks to obtain multilevel information. For example, in Ref. 24, Li. et al. used multiscale multistreaming GCNs, where the multicore parallel temporal convolutional networks obtain more discriminative temporal features. In Ref. 25, the features of different scales were fused selectively. In Ref. 26, multiscale dense connected graph convolution was utilized to enhance local contextual information and obtain flexible temporal graphs. In Ref. 27, multiscale networks and attentional machines were employed to achieve the best recognition results so far. There are now some other methods that can get good results,[28–32] now the one that gets the most attention is the multiscale aggregation scheme and a unified spatial-temporal graph convolutional operator, MS-G3D.[33] In addition, there are also methods that combine GCN and CNN or RNN, such as the attention enhanced graph convolutional LSTM network (AGC-LSTM).[34]

All the methods mentioned above have some improvement in action recognition accuracy, but the computation complexity and the parameter number are also increased. Besides, they do not pay sufficient attention to the subtle differences in similar actions. Compared with existing methods, the proposed STMGCN improves both the nuances of nodes and the ability of spatiotemporal feature extraction. The spatiotemporal excitation network can highlight the spatiotemporal feature information of local nodes. Meanwhile, the STMGCN blocks are fused with the multiscale spatiotemporal network to get increased spatiotemporal features. In such way not only can the recognition of human actions be improved but also the number of model parameters can be reduced as well.

3 Approach

The STMGCN is the stack of these basic blocks, as shown in Fig. 1. There are a total of nine spatiotemporal blocks (B1 to B9) with different parameters which are B1 (3, 64, 1), B2 (64, 64, 1), B3 (64, 64, 1), B4 (64, 128, 1), B5 (128, 128, 1), B6 (128, 128, 1), B7 (128, 256, 1), B8 (256, 256, 1), and B9 (256, 256, 1). The numbers on each block indicate the number of input channels, the number of output channels, and the step size, respectively. Finally, the global average pooling (GAP) and fully connected layers (FC) are performed in turn at the end. The final output is sent to a softmax classifier to obtain the recognition results. The structure diagram of the spatiotemporal module is shown in Fig. 2. The internal structure of each module includes GCN, multiscale temporal convolutional network (MTCN), and spatiotemporal excitation (STE). First, the initial 3D skeleton data $\mathbf{F} \in \mathbf{R}^{(C\times T\times N)}$ is input into the spatial-domain

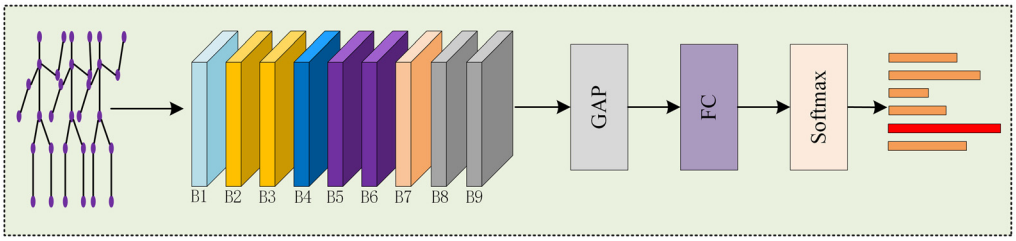

Fig. 1 Framework of STMGCN.

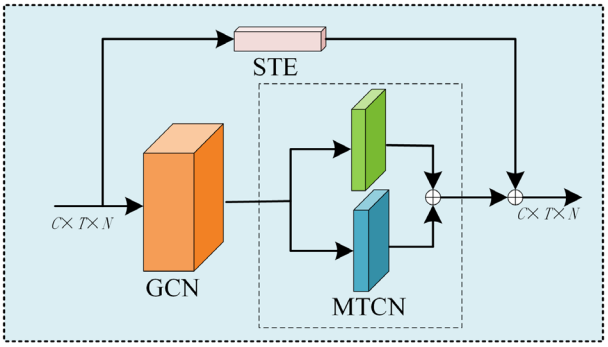

Fig. 2 Illustration of the spatiotemporal block.

GCN to get the spatial feature information of the nodes. After that, the multiscale temporal convolution network is used for the time-domain feature extraction to enhance the temporal correlation between nodes and enrich the temporal difference features between classes. The extracted features are then fused with the key local spatiotemporal features extracted by the STE network. Finally, the fused features are used as the output features $\mathbf{F_o}$ of this network, as shown in the following equation:

$$\mathbf{F}_o = \mathbf{F}_d(\mathbf{f}(\mathbf{F})) + \mathbf{F}_M(\mathbf{F}), \tag{1}$$

where $\mathbf{f}(\cdot)$ is the feature obtained after the graph convolution operation, $\mathbf{F}(\cdot)$ is the feature obtained after the multiscale temporal convolutional network, and $\mathbf{F}_M(\cdot)$ is the feature obtained after STE block.

3.1 *Graph Construction*

The 3D skeleton data for human action recognition can be acquired by motion capture devices such as the Kinect body sensing device or human pose estimation algorithms, such as OpenPose and BlazePose.[35,36] These data are a sequence of frames, and each frame has a set of joint coordinates. Based on the sequence of human joints in the form of given 2D or 3D coordinates, a spatiotemporal skeleton undirected graph G (V, E) with joints as graph nodes and human node structure and natural connectivity of time as graph edges is constructed, as shown in Fig. 3.

The spatiotemporal graph of the skeleton sequence is constructed in two steps. First, within each frame, a spatial skeleton graph is constructed according to the natural skeleton connection relationship of the human body. Second, the edges between frames represent the temporal relationships of the corresponding nodes. In the graph, the node matrix $\mathbf{V} = \{v_{ti} | t = 1,2,\ldots,T, i = 1,2,\ldots,N\}$ includes all the nodes on the skeleton sequence. Here, T is the number of frames, N is the number of joints, and v_{ti} denotes the i'th joint in the t'th frame. The set $\mathbf{E}$ of edges consists of two subsets. One is the connection $\mathbf{E}_S = \{v_{ti}v_{tj}(i,j) \in \mathbf{H}\}$ of nodes within each frame, $\mathbf{H}$ denotes the set of human skeleton joints. The other represents the connection $\mathbf{E}_F = \{v_{ti}v_{(t+1)i}\}$ between different frames.

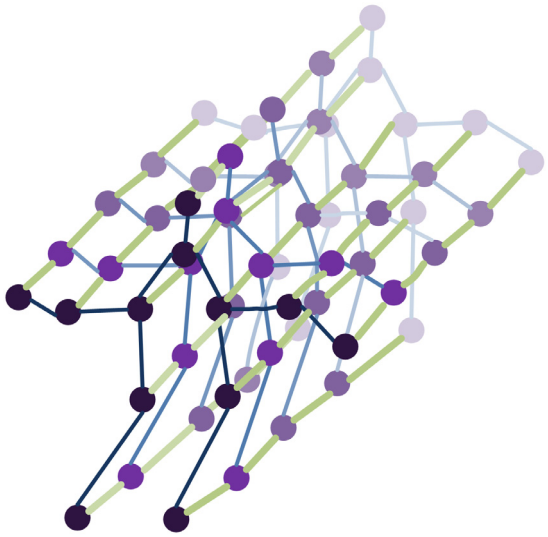

Fig. 3 Illustration of the spatiotemporal graph.

3.1.1 *Spatial graph convolution*

After generating the graph defined above, we perform multilayer spatiotemporal convolution operation on the graph to extract high-level features. The GAP and softmax classifier are used and finally the action category prediction results are obtained.

The mapping strategy is shown in Fig. 4. The topology of the node-space graph within a single frame of the baseline model is represented by the unit matrix $\mathbf{I}$ and the adjacency matrix $\mathbf{A}$. Implementing graph convolution in the spatial dimension is not easy. Specifically, the feature map of the network is a $C \times T \times N$ tensor, where N denotes the number of vertices, T denotes the length of time, and C denotes the number of channels. Assuming that the input features at the n'th layer are $\mathbf{f}^n$ and the output features are $\mathbf{f}^{n+1}$, the expression of the graph convolution is as follows:

$$\mathbf{f}^{n+1} = \sum_{k}^{K_{\max}} \tilde{\mathbf{A}}_k \mathbf{f}^n \mathbf{W}_k \bullet \mathbf{M}_k, \tag{2}$$

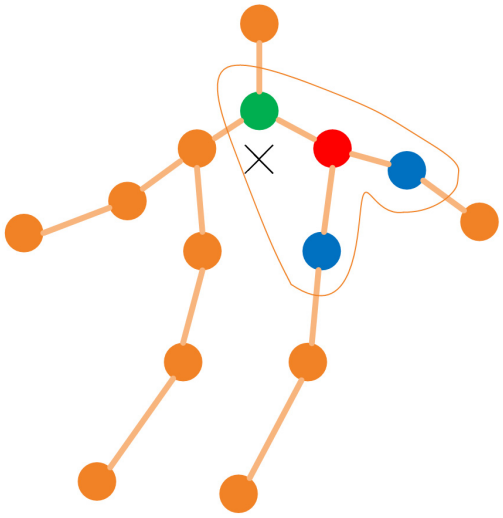

Fig. 4 Illustration of the mapping strategy. Based on the ST-GCN model, the human skeleton is divided into three subsets: the vertex subset (denoted by red nodes), the centripetal subset containing neighboring vertices closer to the center of gravity (denoted by green nodes), and the centrifugal subset containing neighboring vertices farther from the center of gravity (denoted by blue nodes). Here, the symbol $\times$ denotes the gravity center of the human body.[14]

where $\mathbf{W}_k$ is the weight vector of the 1×1 convolution operation, $K_{\max}$ denotes the number of subsets of the adjacency matrix, and $\mathbf{M}_k$ is a learnable matrix that represents the strength of node connectivity:

$$\tilde{\mathbf{A}}_k = \mathbf{D}_k^{-1/2}\mathbf{A}_k\mathbf{D}_k^{-1/2}, \tag{3}$$

$$\mathbf{A} + \mathbf{I} = \sum_k \mathbf{A}_k, \tag{4}$$

where $\mathbf{A}$ denotes the normalized form of the adjacency matrix $\mathbf{A}_k$. The normalized diagonal matrices are that $\mathbf{D}_k^{ii} = \sum_j \mathbf{A}_k^{ij} + l$. Here, $\mathbf{A}_0 = \mathbf{I}$, $\mathbf{A}_1 + \mathbf{A}_2 = \mathbf{A}$, and l is set to 0.001 to avoid empty rows.

3.2 *Multiscale Temporal Convolutional Network*

In terms of the time dimension, it is easy to implement graph convolution operation similar to classical convolution, since the number of neighbors per vertex is fixed to two (corresponding joint in two consecutive frames). Limited by the ability of the convolutional layer in extracting temporal features by the perceptual field, the single-scale temporal convolutional network of the baseline model is difficult to extract the node information with long duration in human action data. In view of this, we use a multiscale temporal convolutional network for the multiscale feature extraction of temporal information, as shown in Fig. 5. The network increases the width of the overall network and enhances its adaptability to the scale on one hand; on the other hand, different branches in the network have different perceptual fields and the scale of time-domain information extraction is different. As training proceeds, the network continuously learns the temporal node features, and the action information expressed by temporal information at different scales can make the network focus more on those salient regions when extracting features.

In Fig. 5, this network performs a channel downscaling on the input data $\mathbf{F} \in \mathbf{R}^{(C \times T \times N)}$ and gets $\mathbf{F} \in \mathbf{R}^{(C/d \times T \times N)}$. Then, the temporal features at different scales can be extracted from the downscaled data. After that, the network dimensionally splices the extracted features and then sums them with those extracted from the residual network. Finally, these obtained features are used as the output of this network $\mathbf{F}_d \in \mathbf{R}^{(C \times T \times N)}$.

The temporal convolutional network in the baseline model performs convolutional operations in time order, in other words, the convolutional operation at moment t occurs only on the data of adjacent frames. The convolutional kernel size taking three, $\mathbf{f}_K$ is a 3×1 convolutional kernel and the input sequence is that $\mathbf{F} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)$. The temporal convolution operation at $\mathbf{x}_T$ can be expressed as

$$\mathbf{F}(x_T) = \mathbf{f}_K \mathbf{x}_{T-K+1}. \tag{5}$$

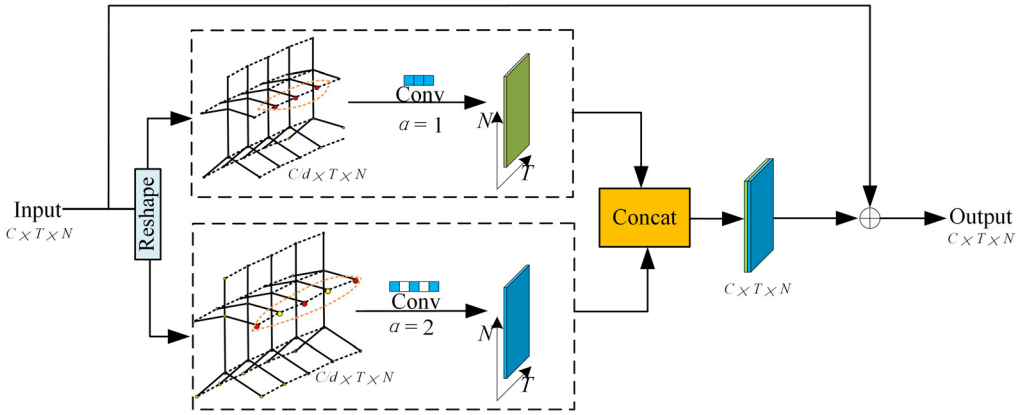

Fig. 5 Structure of multiscale temporal convolutional network.

To enable the temporal network to learn more time-domain features, the model should be able to utilize more frames of the action data. In this paper, the multiscale temporal feature information is extracted by a dilated convolutional method, and the residual connectivity is also utilized to facilitate training, expressed as follows:

$$\mathbf{F}_d(x_T) = \sum_{\alpha=1}^{d} \mathbf{f}_{\mathrm{K}} \mathbf{x}_{T-\mathrm{K}+1} + \mathbf{X}, \tag{6}$$

where α is the dilation rate of the dilation convolution, d denotes the scale size, and $\mathbf{X}$ is the result of the input data $\mathbf{F} \in \mathbf{R}^{(C\times T\times N)}$ after 1×1 convolutional mapping, $\mathbf{X} \in \mathbf{R}^{(C\times T\times N)}$.

The multiscale temporal convolutional network keeps the resolution of the output feature map constant by utilizing the dilation convolution. The dilation rate α of the dilation convolution takes the values $[1,2,\ldots,d]$. When $\alpha = 1$, the perceptual field of the convolution kernel is three and the parameter number is three. When $\alpha = 2$, the perceptual field of the convolution kernel is five, but the parameter number remains unchanged, as shown in Fig. 6.

Compared with the baseline model, the multiscale temporal convolutional network proposed in this paper has the ability of perceiving multiple scale features in time-domain and achieves better recognition results with fewer parameters.

3.3 *Spatiotemporal Excitation Network*

Complex human actions are not only related to space but also often time-dependent. In view of this, we need to enhance the joint spatiotemporal feature extraction ability to improve the action recognition effect. The baseline model cannot sufficiently extract the spatiotemporal feature information of the skeleton nodes when recognizing human actions, however. Given this, we extend the spatiotemporal excitation network based on video data to the human skeleton data and propose a spatiotemporal excitation network that can effectively model dynamic skeleton data. Since the method proposed in Ref. 3 is for action recognition of video data, the data are two-dimensional (2D) and contain the adjacency matrix without the interframe depth of the actions in the video. Unlike the method in Ref. 3 which uses 3D convolution to get the spatiotemporal information in action videos, the spatiotemporal feature map in this paper is compressed into a single channel after channel pooling. Therefore, the spatiotemporal feature map is transformed into the 2D data, and then 2D convolution is used to excite the important spatiotemporal information. The network focuses on the dynamic feature information of spatiotemporal domain between local nodes of the human body, which effectively enhances the role of key nodes in similar actions and facilitates the recognition of similar actions. The structure of the spatiotemporal excitation network is given in the STE network block in Fig. 7.

Specifically, given $\mathbf{F} \in \mathbf{R}^{(C\times T\times N)}$, the global spatiotemporal tensor $\mathbf{F}_1 \in \mathbf{R}^{(1\times T\times N)}$ is first obtained by channel pooling of the input data $\mathbf{F}$. The 2D convolution is performed on the pooled data $\mathbf{F}_1$ with a 3×3 convolution kernel $\mathbf{K}_1$, formulated as follows:

$$\mathbf{F}_r = \mathbf{K}_1 \mathbf{F}_1. \tag{7}$$

By convolving the obtained features $\mathbf{F}_r$ we obtain a spatiotemporal feature mask $\mathbf{M}$, which is equivalent to the spatiotemporal attentional feature map:

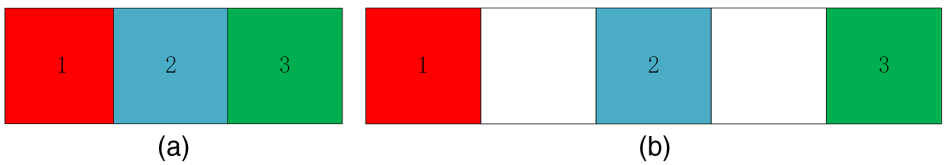

Fig. 6 Convolution kernels with different dilation rates. (a) Convolution kernel 3*1 $\alpha = 1$ and (b) convolution kernel 3*1 $\alpha = 2$.

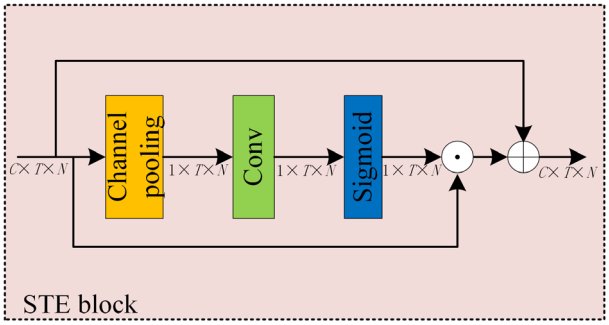

Fig. 7 Illustration of the STE block.

$$\mathbf{M} = \delta(\mathbf{F}_r). \tag{8}$$

After dot product operation between $\mathbf{M}$ and initial input tensor $\mathbf{F}$, the result is added with $\mathbf{F}$ to get $\mathbf{F}_M$. Here, by dot product operation, the excitation of key spatiotemporal features from the input features $\mathbf{F}$ with the template $\mathbf{M}$ can be realized:

$$\mathbf{F}_M = \mathbf{F} + \mathbf{F} \bullet \mathbf{M}. \tag{9}$$

The spatiotemporal excitation network is computationally more efficient than the conventional 2D convolutional operation, for the order of the parameter matrix and the number of parameters are effectively reduced through channel pooling. Each channel of $\mathbf{F}$ can perceive important spatiotemporal information through the activation function $\delta(\mathbf{F}_r)$ without increasing the number of parameters. In the overall network structure, the spatiotemporal excitation network replaces the residual block in the baseline model, which not only increases the network depth but also makes up for the lack of spatiotemporal feature extraction ability of the baseline model. Moreover, it enhances the role of key nodes, improving the recognition effect on similar actions consequently.

4 Tests

4.1 *Datasets*

NTU-RGB+D60 is a publicly available 3D human skeleton action dataset containing 60 action categories with a total of 56,880 video samples, performed by 40 volunteers in the age range of 10 to 35. The action videos of each category were captured by three cameras, and all of them were filmed at the same chosen height but at different horizontal angles, that is, −45 deg, 0 deg, and 45 deg, respectively. The dataset uses the Kinect depth sensor to detect a sequence of 3D skeletal points for each frame of the human body. Each subject has 25 joints in the skeleton sequences and no more than two persons in each video. Two partitioning methods, X-Sub and X-View, are provided in Ref. 18 for dataset partitioning. X-Sub divides the dataset into a training set (40,320 videos) and a validation set (16,560 videos) by the participants in the captured videos; X-View uses 37,920 videos captured by cameras 1 and 3 as the training set, while 18,960 videos from camera 2 as the validation set.

NTU-RGB+D120 is an extension of the NTU-RGB+D60 dataset. About 106 participants completed 120 action categories, yielding 113,945 action samples with 32 different camera devices and 155 viewpoints. Different from NTU-RGB+D60, NTU-RGB+D120 is divided by X-Sub and X-Set in Ref. 19. By X-Sub, the dataset is divided into a training set (63,026 videos) and a validation set (50,919 videos) according to the participants in the captured videos. By X-Set, the dataset is divided according to the video numbers, the videos with even numbers as the training set (54,468 videos) and the videos with odd numbers as the test set (59,477 videos).

Kinetics 400 is a large-scale and high-quality dataset of YouTube video websites. It contains over 300,000 video clips in 400 categories. The skeleton dataset is extracted by the OpenPose

toolbox, and each skeleton map contains 18 human joints, representing each joint in 2D spatial coordinates and predicted confidence scores.[20]

Pytorch is employed as a deep learning framework for all the tests in this paper.[37] The weight decay rate is set to 0.0005, the epoch is set to 70, the initial learning rate is set to 0.05, and batchsize is 32.

Inspired by the two-stream method in baseline model 2s-AGCN, we used the same two-stream fusion strategy. Two streams, namely joint and bone, are used. The joint stream uses the original skeleton coordinates as input, and the bone stream uses the differential of spatial coordinates as input. Each stream was separately input into the network to get the softmax scores. Finally, the maximum scores of the two streams were added to obtain the fused score.

4.2 *Ablation Tests*

To verify the proposed method, we designed several ablation comparison tests, and 2s-AGCN was selected as the baseline model. Three groups of tests were done and performance comparisons were conducted among several widely applied models from different aspects.

To verify the effectiveness of multiscale time-domain convolutional networks in capturing complex time-domain features, the tests of group 1 were designed and conducted to evaluate the impact of the key parameter d. The parameter d can be used to characterize the strength of multiscale temporal feature extraction, representing the range of values of the expansion rate. In this paper, some tests were conducted to get the optimal value of the parameter d under X-Sub division. Here, d was taken in [1,2,3,4] sequentially. The test results are listed in Table 1. It can be seen from Table 1 that the best result is obtained at $d = 2$ with the accuracy of 89.6%, followed by $d = 3$, while the worst result is obtained when $d = 1$. As can be seen, the appropriate dilation rate is very important, and the excessive dilation rate may lose the continuity of information and affect the recognition results. The number of parameters was also compared in this test. The number of parameters at $d = 2$ is 1.6M, about 1.9M lower compared with that of the baseline model, which demonstrates its advantage in calculation amount.

The effectiveness of the MTCN and STE network was evaluated on the NTU-RGB+D60 dataset, and the results are listed in Table 2. Compared with the baseline model, for X-Sub division and X-View division, MTCN achieved the improvement of 1.1% and 0.3%, and STE network achieved the improvement of 1.2% and 0.6%, respectively. This fully demonstrates the superiority of these two methods.

4.3 *Visualization Analysis*

As shown in Fig. 8, the left side is the adjacency matrix of the layer 8 learned from the baseline network, and the right side corresponds to the matrix learned from the network proposed in this paper. The grayscale of each element in the matrix indicates the strength of the connection. As can be seen, the values of the adjacency matrix learned by the baseline model tend to average out, making it difficult to distinguish the connection characteristics of nodes with similar actions. Whereas the method in this paper reduces the connection strength of nodes that are not related

Table 1 Comparison of accuracy and parameter quantities at different scales.

Method	Parameter quantities (M)	Accuracy under X-Sub (%)
Baseline	3.5	88.5
Baseline + MTCN ($d = 1$)	1.6	89.2
Baseline +MTCN ($d = 2$)	**1.6**	**89.6**
Baseline +MTCN ($d = 3$)	1.6	89.5
Baseline +MTCN ($d = 4$)	1.6	89.4

Note: Our method and its best recognition accuracy are shown in bold.

Table 2 Test results of MTCN network and STE network on NTU-RGB+D60 dataset.

Method	Accuracy (%)	
	X-Sub	X-View
Baseline (joint)	—	93.7
Baseline (bone)	—	93.2
Baseline (J+B)	88.5	95.1
Baseline +MTCN (joint)	87.4	94.3
Baseline +MTCN (bone)	87.9	94.2
Baseline + MTCN (J +B)	**89.6**	**95.4**
Baseline +STE (joint)	87.3	94.6
Baseline + STE (bone)	88.6	94.2
Baseline + STE (J+B)	**89.7**	**95.7**

Note: Our method and its best recognition accuracy are shown in bold.

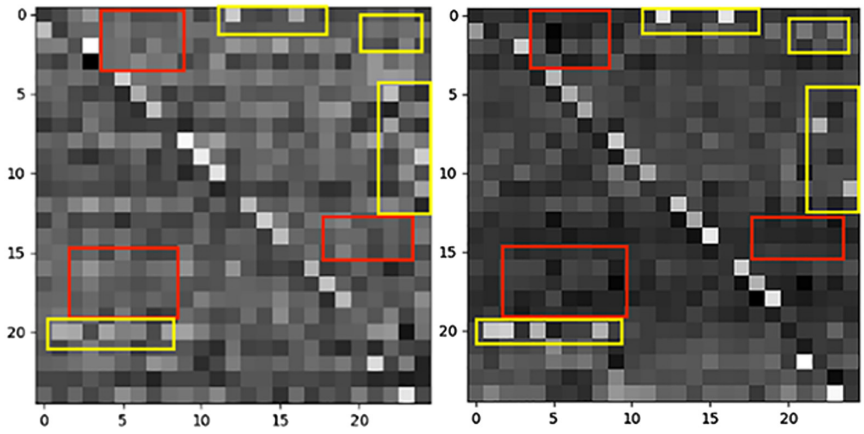

Fig. 8 Example of the learned adjacency matrix.

to the action (the red boxes in Fig 8), such as nodes 14, 15, 16, and 18 that are not related to the reading action. In addition, it effectively enhances the connection strength of important related nodes in the action (the yellow boxes in Fig 8), such as nodes 21, 22, and 24, which are related to the reading action.

A reading action sample sequence is visualized in a feature map, as shown in Fig. 9. The vertical coordinates of the feature map indicate the node number and the horizontal coordinates

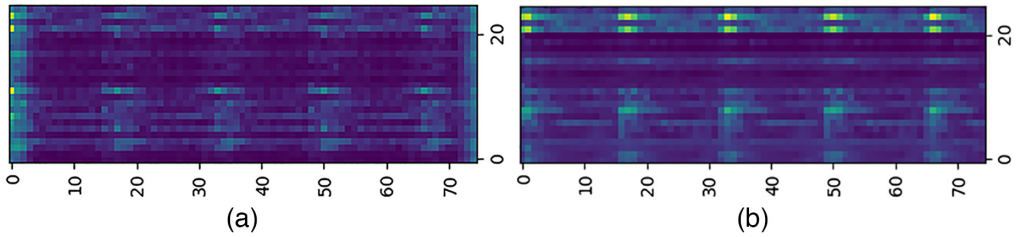

Fig. 9 Visualization of feature maps for a reading action sample. (a) Feature map of the baseline model and (b) feature map of our method.

indicate the frame sequence number. Figures 9(a) and 9(b) show the feature map in the layer 8 learned by the baseline network and the proposed network, respectively. And Fig. 10 shows the feature visualization with key points for a selected frame of the above feature map. It can be found by comparison that the proposed method can significantly enhance the connection information of key nodes in the action in the time-domain, which fully demonstrates the superiority of the method in this paper.

In Fig. 11, examples of skeleton-based action samples are presented on the NTU-RGB +D60. For each action, five frames are selected in the chronological order to represent the whole action. The one on the left is the action frame for reading, and the right is writing. It can be seen that spatial features of the action of single frames are very similar, but the temporal features of "writing" versus "reading" vary more significantly. In this way, our view that the extraction of temporal difference features will increase the recognition effect of similar actions is thus verified.

4.4 *Results and Comparative Analysis*

To explore the effectiveness of STMGCNs for human action recognition, we performed corresponding tests on the public datasets NTU-RGB+D60 and NTU-RGB+D120, and the test results are listed in Tables 3 and 4, respectively. In terms of recognition accuracy, the performance effect of our method is better than that of most models.

Compared with the baseline model, the accuracy of the proposed method was improved by 1.7% and 0.4% under X-Sub and X-View division on the NTU-RGB+D60 dataset. The accuracy increment of the proposed method was 8.7% and 1.0% under X-Sub, and 7.2% and 0.5% under X-View, compared with that of the typical methods ST-GCN[23] and AGC-LSTM.[33] On the NTU-RGB+D120 dataset, the accuracy of the proposed method was improved by 3.2% and

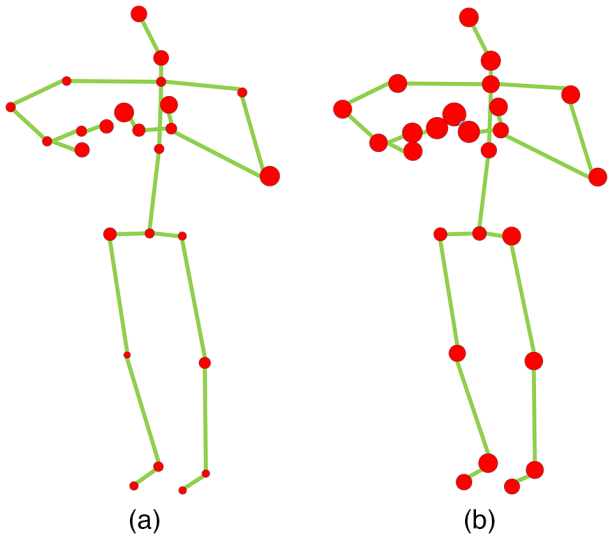

Fig. 10 Visualization of output features. Each dot represents one joint, and its size represents the degree of the attention to the node. The green lines represent the physical connections of human body. (a) Output features of the baseline model and (b) output features of our method.

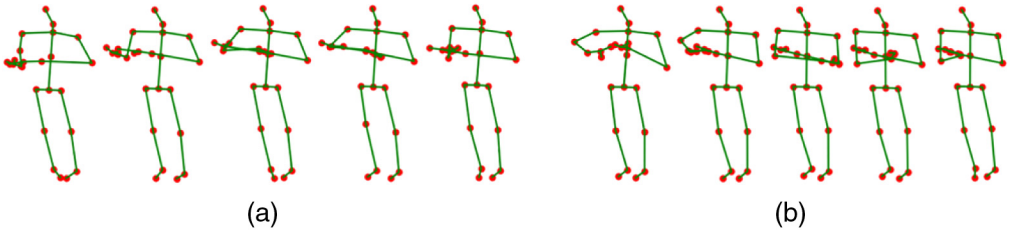

Fig. 11 Visualization of skeleton-based action samples. (a) Reading and (b) writing.

Table 3 Comparison of test results on NTU-RGB+D60 dataset.

Method	Accuracy (%)	
	X-Sub	X-View
IndRNN[6]	81.8	88.0
TS-LSTM[8]	74.6	81.3
JTM[10]	76.3	80.8
GA spatiotemporal[12]	82.8	90.0
AGC-LSTM[34]	89.2	95.0
ST-GCN[23]	81.5	88.3
AS-GCN[16]	86.8	94.2
PB-GCN[15]	87.5	93.2
MSGCN[24]	88.8	95.7
MS-G3D[33]	91.5	96.2
STA-GCN[27]	92.4	96.5
2s-AGCN[14]	88.5	95.1
Our STMGCN (joint)	88.2	94.0
Our STMGCN (bone)	88.3	94.4
Our STMGCN	**90.2**	**95.5**

Note: Our method and its best recognition accuracy are shown in bold.

Table 4 Comparison of test results on NTU-RGB+D120 dataset.

Method	Accuracy (%)	
	X-Sub	X-Set
2s-AGCN[14]	82.9	84.9
SGN[17]	79.2	81.5
AS-GCN[16]	77.7	78.9
MS-G3D[33]	86.9	88.4
Our STMGCN	**86.1**	**87.0**

Note: Our method and its best recognition accuracy are shown in bold.

2.1% under X-Sub and X-Set, respectively, compared with the baseline model. It is obvious that the proposed method outperforms most methods. Compared with the latest MS-G3D model, although the recognition accuracy of the model is slightly lower, the total number of parameters of the proposed model is only half that of the MS-G3D model.

In the Kinetics-Skeleton dataset, our method was compared with several of the most classical methods in Table 5. Following the evaluation method in 2s-AGCN, we trained the model in the training set and provided the accuracy of top-1 and top-5 on the validation set. We see that our method improves of 0.9% for top-1 and 1.4% for top-5 compared with 2s-AGCN, demonstrating the effectiveness of our method.

Table 5 Comparison of test results on Kinetics-Skeleton dataset.

Method	Accuracy (%)	
	Top-1	Top-5
ST-GCN[23]	30.7	52.8
AS-GCN[16]	34.8	56.5
2s-AGCN[14]	36.1	58.7
MS-G3D[33]	38.0	60.9
Our STMGCN	**37.0**	**60.1**

Note: Our method and its best recognition accuracy are shown in bold.

On the NTU-RGB+D60 dataset, we compared the proposed method with the baseline model on each category adopting X-Sub, and the results are shown in Fig. 12. As can be seen, the proposed method outperforms the baseline in most cases and even improves the accuracy by more than 5% on categories 9, 10, 11, 12, 15, 29, 30, 32, and 52. It is owing to the better feature extraction ability of our proposed method for local subtle movement in human action. More useful information in the temporal feature extraction process can be retained owing to the usage of the multiscale temporal convolutional network which extracts features of time-domain information at different scales, and cascades and fuses the time-domain features at different scales.

Compared with the latest models of STA-GCN and MS-G3D, the recognition accuracy of our method is slightly lower, but our method has shown superior results in the vast majority of similar actions, as shown in Fig. 13. For the similar actions, such as clapping, reading, writing, tearing up paper, putting on a shoe, and taking off a shoe, they differ by small changes in arm movements, thus creating challenges for skeleton-based action recognition. For these similar actions, the accuracy of our method has improved 7.2% compared with that of the baseline model, 3.5% compared with that of the MS-G3D, and 4% compared with that of the STA-GCN on the NTU-RGB+D60 dataset X-Sub. This demonstrates that our method significantly increases the recognition of similar actions by increasing the temporal feature extraction capability and the importance of important nodes.

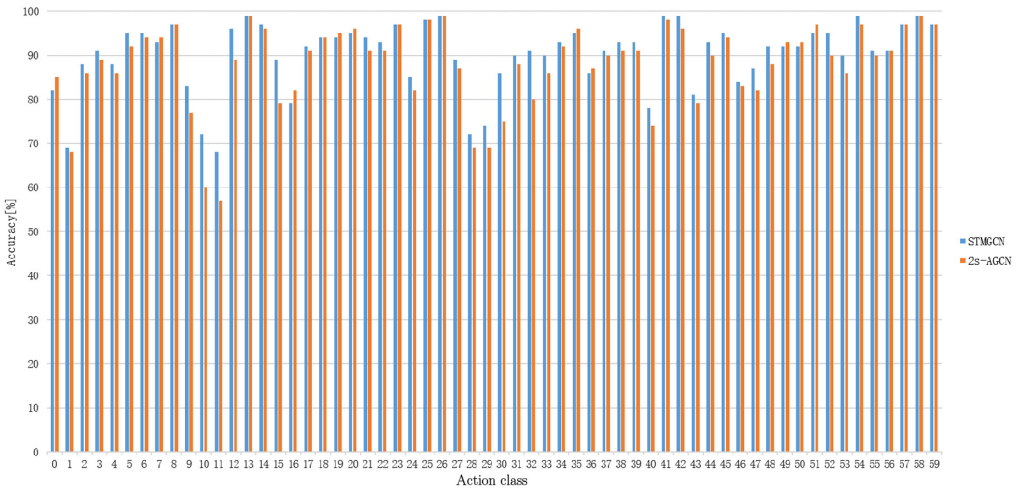

Fig. 12 Accuracy comparison between the proposed method and 2s-AGCN based on X-Sub evaluation. The horizontal coordinate is the serial number of the action category, and the vertical coordinate is the recognition accuracy. The specific action names are referred to the official website of the public dataset NTU-RGB+D.

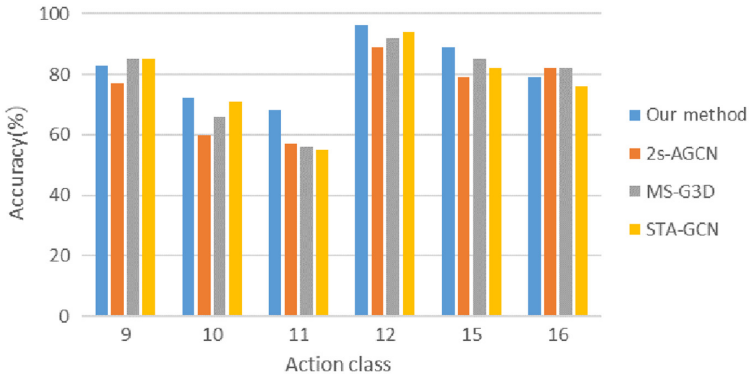

Fig. 13 Comparison of the performance on similar classes.

5 Conclusion

In this paper, we propose a new model for human action recognition based on human skeleton data. It consists of a spatio graph convolutional network, a multiscale temporal convolutional network, and a spatiotemporal excitation network. The multiscale temporal convolutional networks extract the time-domain features from the data at different scales, increasing the overall width of the network and enabling more focus on significant regions during feature extraction. The spatiotemporal excitation networks with channel pooling and 2D convolution operations can activate the critical spatiotemporal information and enhance the role of key nodes in similar actions, thus improving the action recognition effect. The proposed model was tested on three large-scale action recognition datasets, NTU-RGB+D60, NTU-RGB+D120, and Kinetics-Skeleton. The STMGCN model achieves considerable improvement compared with the baseline model 2s-AGCN and other models in recognition accuracy of similar human actions. In future, we plan to fuse the RGB, IR, and human skeleton data to make full use of multimodal information to improve the accuracy of human action recognition.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (Grant No. 62076093) and S&T Program of Hebei (Grant No. SZX2020034).

References

1. Y. Zhu et al., "A review of human action recognition based on deep learning," *Acta Autom. Sin.* **42**, 848–857 (2016).
2. S. Zennaro et al., "Performance evaluation of the 1st and 2nd generation Kinect for multimedia applications," in *IEEE Int. Conf. Multimedia and Expo (ICME)*, IEEE, Turin, pp. 1–6 (2015).
3. Z. Wang, Q. She, and A. Smolic, "ACTION-Net: multipath excitation for action recognition," in *IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, IEEE, Nashville, Tennessee, pp. 13209–13218 (2021).
4. Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Networks* **5**(2), 157–166 (1994).
5. G. Chéron, I. Laptev, and C. Schmid, "P-CNN: pose-based CNN features for action recognition," in *IEEE Int. Conf. Comput. Vision*, IEEE, Santiago, pp. 3218–3226 (2015).
6. S. Li et al., "Independently recurrent neural network (IndRNN): building a longer and deeper RNN," in *IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, IEEE, Salt Lake City, Utah, pp. 5457–5466 (2018).
7. J. Liu et al., "Global context-aware attention LSTM networks for 3D action recognition," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, IEEE, Honolulu, Hawaii, pp. 3671–3680 (2017).

8. I. Lee et al., "Ensemble deep learning for skeleton-based action recognition using temporal sliding LSTM networks," in *IEEE Int. Conf. Comput. Vision*, IEEE, Venice, pp. 1012–1020 (2017).
9. W. Zheng et al., "Relational network for skeleton-based action recognition," in *IEEE Int. Conf. Multimedia and Expo*, IEEE, Shanghai, pp. 826–831 (2019).
10. P. C. Wang et al., "Action recognition based on joint trajectory maps with convolutional neural networks," *Knowl.-Based Syst.* **158**, 43–53 (2018).
11. C. Caetano et al., "SkeleMotion: a new representation of skeleton joint sequences based on motion information for 3D action recognition," in *16th IEEE Int. Conf. Adv. Video and Signal Based Surveill.*, IEEE, Taipei, pp. 1–8 (2019).
12. Y. Li et al. "Learning shape-motion representations from geometric algebra spatio-temporal model for skeleton-based action recognition," in *IEEE Int. Conf. Multimedia and Expo*, IEEE, Shanghai, pp. 1066–1071 (2019).
13. T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *5th Int. Conf. Learn. Represent.*, OpenReview.net, Toulon, France (2017).
14. L. Shi et al., "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, IEEE, Long Beach, California, pp. 12018–12027 (2019).
15. K. C. Thakkar and P. J. Narayanan. "Part-based graph convolutional network for action recognition," in *Brit. Mach. Vision Conf. 2018*, BMVC 2018: 270, Newcastle, UK (2018).
16. M. Li et al., "Actional-structural graph convolutional networks for skeleton-based action recognition," in *IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, IEEE, Long Beach, California, pp. 3590–3598 (2019).
17. P. Zhang et al., "Semantics-guided neural networks for efficient skeleton-based human action recognition," in *IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, IEEE, Seattle, Washington, pp. 1109–1118 (2020).
18. A. Shahroudy et al., "NTU RGB+D: a large scale dataset for 3D human activity analysis," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, IEEE, Las Vegas, Nevada, pp. 1010–1019 (2016).
19. J. Liu et al., "NTU RGB+D 120: a large-scale benchmark for 3D human activity understanding," *IEEE Trans. Pattern Anal. Mach. Intell.* **42**(10), 2684–2701 (2020).
20. J. Carreira and A. Zisserman, "Quo Vadis, action recognition? A new model and the kinetics dataset," in *IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)*, IEEE, Honolulu, Hawaii, pp. 4724–4733 (2017).
21. S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.* **9**(8), 1735–1780 (1997).
22. K. Cho et al., "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Conf. Empirical Methods in Nat. Lang. Process.*, Association for Computational Linguistics, Doha, Qatar, pp. 1724–1734 (2014).
23. S. J. Yan, Y. J. Xiong, and D. H. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. Thirty-Second AAAI Conf. Artif. Intell. and Thirtieth Innov. Appl. of Artif. Intell. Conf. and Eighth AAAI Symp. on Educ. Adv. in Artif. Intell.*, AAAI Press, New Orleans, Louisiana, Article 912, pp. 7444–7452 (2018).
24. W. Li et al., "Skeleton-based action recognition using multi-scale and multi-stream improved graph convolutional network," *IEEE Access* **8**, 144529–144542 (2020).
25. Y. Fan et al., "Multi-scale adaptive graph convolutional network for skeleton-based action recognition," in *15th Int. Conf. Comput. Sci. & Educ. (ICCSE)*, Delft, The Netherlands, pp. 517–522 (2020).
26. H. Xia and X. Gao, "Multi-scale mixed dense graph convolution network for skeleton-based action recognition," *IEEE Access* **9**, 36475–36484 (2021).
27. M. C. Le, *Skeleton-Based Human Action Recognition Using Spatio-Temporal Attention Graph Convolutional Networks*, Lappeenranta-Lahti University of Technology, Finland (2022).
28. F. F. Ye et al., "Dynamic GCN: context-enriched topology learning for skeleton-based action recognition," in *Proc. 28th ACM Int. Conf. Multimedia*, Association for Computing Machinery, New York, pp. 55–63 (2020).

29. L. Shi et al., "Skeleton-based action recognition with multi-stream adaptive graph convolutional networks," *IEEE Trans. Image Process.* **29**, 9532–9545 (2020).
30. X. Zhang, C. Xu, and D. Tao, "Context aware graph convolution for skeleton-based action recognition," in *IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, IEEE, Seattle, Washington, pp. 14321–14330 (2020).
31. Y.-F. Song, Z. Zhang, and L. Wang, "Richly activated graph convolutional network for action recognition with incomplete skeletons," in *IEEE Int. Conf. Image Process. (ICIP)*, IEEE, Taipei, pp. 1–5 (2019).
32. H. Yang et al., "Feedback graph convolutional network for skeleton-based action recognition," *IEEE Trans. Image Process.* **31**, 164–175 (2022).
33. Z. Liu et al., "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, IEEE, Seattle, Washington, pp. 140–149 (2020).
34. C. Si et al., "An attention enhanced graph convolutional LSTM network for skeleton-based action recognition," in *IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, IEEE, Long Beach, California, pp. 1227–1236 (2019).
35. Z. Cao et al., "OpenPose: realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.* **43**(1), 172–186 (2021).
36. V. Bazarevsky et al., "BlazePose: on-device real-time body pose tracking," in *CVPR Workshop on Comput. Vision for Augmented and Virtual Reality*, Seattle, Washington, USA (2020).
37. A. Paszke et al., "Automatic differentiation in PyTorch," in *Neural Inf. Process. Syst.* (2017).

Yincheng Qi received his BS, MS, and PhD degrees from North China Electric Power University, Baoding, China, in 1990, 1998, and 2009, respectively. He is currently a professor with North China Electric Power University. His research interests include electric power system communication and information processing, and computer vision in electric power systems. He is the vice chairman of Electric Power Communication Study Committee for CSEE.

Baoli Wang received her BS degree in communication engineering from North China Electric Power University, Baoding, China, in 2021. She is currently pursuing her MS degree in information and communication engineering at North China Electric Power University, Baoding, China. Her research interest is human action recognition.

Boqiang Shi received his BS degree in electronic information engineering from Hebei University of Economics and Business, Shijiazhuang, China, in 2018. And he received his MS degree in electronics and communication engineering from North China Electric Power University, Baoding, China, in 2022. His research interest is human action recognition.

Ke Zhang is a professor at North China Electric Power University, Baoding, China. He received his ME degree in signal and information processing from North China Electric Power University, Baoding, China, in 2006, and his PhD in signal and information processing from Beijing University of Posts and Telecommunications, Beijing, China, in 2012. His research interests include computer vision, deep learning, machine learning, robot navigation, natural language processing, and spatial relation description.