Applied Intelligence (2023) 53:17796–17808
https://doi.org/10.1007/s10489-022-04442-y

Spatial adaptive graph convolutional network for skeleton-based action recognition

Qilin Zhu[1] · Hongmin Deng[1]

Accepted: 28 December 2022 / Published online: 13 January 2023

Abstract

In recent years, great achievements have been made in graph convolutional network (GCN) for non-Euclidean spatial data feature extraction, especially the skeleton-based feature extraction. However, the fixed graph structure determined by the fixed adjacency matrix usually causes the problems such as the weak spatial modeling ability, the unsatisfactory generalization performance, the excessively large number of model parameters, and so on. In this paper, a spatially adaptive residual graph convolutional network (SARGCN) is proposed for action recognition based on skeleton feature extraction. Firstly, the uniform and fixed topology is not required in our graph. Secondly, a learnable parameter matrix is added to the GCN operation, which can enhance the model's capabilities of feature extraction and generalization, while reducing the number of parameters. Therefore, compared with the several existing models mentioned in this paper, the least number of parameters are used in our model while ensuring the comparable recognition accuracy. Finally, inspired by the ResNet architecture, a residual connection is introduced in GCN to obtain higher accuracy at lower computational costs and learning difficulties. Extensive experimental on two large-scale datasets results validate the effectiveness of our proposed approach, namely NTU RGB+D 60 and NTU RGB+D 120.

Keywords Action recognition · GCN · SARGCN · Skeleton

1 Introduction

With the development of deep learning and its great potential applications in the field of computer vision, skeleton-based action recognition by using deep learning has drawn widespread attention. For example, skeleton-based action recognition technology has been widely used in the fields of video surveillance, human-computer interaction, and video understanding. Compared with the traditional methods based on two-dimensional RGB data [1, 2], the skeleton-based methods are more adaptable to complex dynamic environments, and can fully depict the spatial and temporal dynamics of human behavior with high robustness and computational efficiency [3–5]. In addition, a natural topological structure in non-Euclidean space can be constructed based on the skeleton data, with vertices obtained from the joints and edges from the connections between the joints, which can show the posture of the human body more concisely and avoid the interference of complex environmental factors [6].

Here we shall first give a brief introduction to some deep learning methods including RNN-based methods and CNN-based methods.

RNN-based methods: In some previous studies, some researchers applied RNN to the feature extraction of human skeleton sequences. In the RNN-based methods, the skeleton data are usually modeled as a sequence of coordinate vectors in the spatial and temporal dimensions, where each vector represents a human joint [3, 5, 7–10]. Song et al. [3] proposed a framework for learning the spatiotemporal features of skeleton data using an attention mechanism. In those attention subnets of spatial and temporal dimensions, the authors used the LSTM networks to learn the relationship of the nodes between the current frame and the previous frame. In [5], Du et al. proposed an end-to-end hierarchical RNN for skeleton-based action recognition, dividing the human skeleton data

✉ Hongmin Deng
hm_deng@scu.edu.cn

Qilin Zhu
zhuqilin@stu.scu.edu.cn

[1] College of Electronics and Information Engineering, Sichuan University, Chengdu, 610065, Sichuan, China

into five parts, and then feeding them into five Bi-RNN sub-networks for the feature extraction. Zhang et al. [8] introduced a view self-adjusting scheme based on the LSTM mechanism to dynamically recognize the actions from skeleton data. In [10], An end-to-end fully connected (FC) deep LSTM network was proposed for skeleton-based action recognition, and a new dropout algorithm and new regularization method were introduced to train the network and extract co-occurrence features of skeleton data.

CNN-based methods: Some previous researchers have made good achievements in applying CNN to feature extraction for skeleton-based action recognition [11–15]. Kim and Reiter [11] used the spatiotemporal information of the skeleton sequence and proposed a new model called temporal convolutional neural network (TCN) for 3D human action recognition, which provided interpretable spatiotemporal representation for learning and training explicitly. Li et al. [12] constructed three views in the spatial domain and made full use of the temporal and spatial information to extract features, where the recognition scores from all views can be combined by a multiple fusion method. Ke et al. [13] proposed a multi-task learning network (MTLN), where the skeleton sequences with arbitrary length were first transformed into three clips, then these clips were fed into a deep CNN model for feature extraction and action recognition.

However, the above-mentioned RNN-based methods and CNN-based methods are not very effective in the action recognition of skeleton data because it is difficult for RNN and CNN to represent the topology of skeleton data.

It was the first paper that presented the graph convolutional neural network to solve the problem of action recognition based on the human skeleton [16]. Since then, more and more researchers have been devoting themselves to the field of human action recognition by using GCNs, and many kinds of GCNs have been proposed such as 2s-AGCN, AS-GCN, ResGCN, MS-AAGCN, and so on.

Compared with the typical deep neural networks, significant performance improvements in recognition accuracy have been achieved by using GCNs. However, some large challenges for the GCN methods are as follows: 1) Existing GCN models are not enough to adaptively extract the general skeletal spatial features; 2) The existing GCN models, with relatively large architectures and more parameters, are difficult to be trained quickly and accurately.

In response to the above problems, in this article we are aimed at improving the spatial feature extraction on the skeleton sequence while taking into account the varying degrees of the significance of various skeleton joints and their connections in various actions, so a learnable parameter matrix is added to enhance the model's extraction of spatial features when training. We introduce a six-layer GCN structure for spatial feature extraction by reducing the scale and complexity of the model and adding residual connections [17] to avoid the performance degradation of the model. Experiments on the NTU RGB+D 60 [18] and NTU RGB+D 120 [19] datasets show the good performance of our proposed model.

The highlights of this paper are mainly reflected in the following two aspects:

1. A novel six-layer spatially adaptive residual graph convolutional network (SARGCN) is proposed while enhancing the model's capability of spatial feature extraction from the skeleton data.
2. Comparable performance in recognition accuracy has been achieved at the minimum number of parameters and much lower computation costs through a lot of comparison experiments, in contrast to the several existing models mentioned in this paper.

2 Related work

Recently, the graph convolutional neural network, which extends the convolution neural network from image recognition to graph recognition, has been widely applied in many fields [20–23]. Related work is reviewed briefly in this section, including the GCN-based attention mechanism and human action recognition.

2.1 Skeleton-based action recognition for GCN methods

As mentioned in Section 1, in the inspiration of Yan and his group's founding work, many researchers started to study action recognition based on skeleton sequences [6, 24–27]. Shi et al. [6] proposed an adaptive graph convolutional network (AGCN) structure with a better topology learning ability for different graph convolutional layers and end-to-end skeleton samples, and it could also be better suited for recognition tasks and its hierarchical structure. At the same time, the second-order information was combined with the first-order information through the dual-stream structure, which played a good role in improving the performance of the model. They further improved the 2s-AGCN to be a multi-stream structure called MS-AAGCN, and added attention mechanism into this new model [26]. However, this multi-stream structure showed well-performance at the cost of high computational complexity and a large number of model parameters. Li et al. [24] proposed an A-link inference model (AIM) to infer actional links that could capture the potential relationships of specific actions, and also proposed an action-structure graph convolutional network (AS-GCN) based on multiple graphs to extract useful space and time information. Song et al. [25] proposed an effective but robust baseline model based on GCN, which integrated the multiple input branches module and partial attention block into the residual graph convolutional network with bottleneck structure.

2.2 Attention mechanism

Attention module has become an important concept in neural networks and has been fully studied in different application fields, e.g. action recognition, target detection, natural language processing, and so on. Baradel et al. [28] proposed a new human action recognition mechanism based on spatiotemporal attention of human pose. Song et al. [3] proposed a spatiotemporal attention model based on LSTM, which could automatically learn the importance levels of different nodes and different frames, and give different attention weights to each frame and node. Si et al. [29] introduced an attention mechanism based on LSTM to enhance the features of key nodes, which helped to improve spatiotemporal representation. In the above three models, the attention mechanism was separately employed to each frame of the skeleton sequence. In addition, the traditional attention module was usually realized through multi-layer perception, without considering the overall relevance of spatiotemporal attention.

3 The proposed approach

In this section, firstly, the existing works on how to apply spatiotemporal graph convolutional network to human action recognition based on skeleton data are reviewed. Then the implementation of our proposed SARGCN model is elaborated.

3.1 The principles of ST-GCN

The skeleton data in the video can be captured by depth camera equipment or algorithms. Generally, this data is a sequence of video frames, with each frame having a combination of 2D/3D joint coordinates. The skeleton sequence is actually composed of a 3D tensor whose shape is $C \times T \times N$, which means that there are C channels, T frames, and N nodes. Simultaneously, an undirected spatiotemporal graph $G = (V, E)$ is constructed on a skeleton sequence with N joints and T frames, where $V = \left\{v_i^t \mid i = 1, 2, ..., N; t = 1, 2, ..., T\right\}$ represents the set of all joints, and E represents the set of connecting edges. The edge set E consists of two parts: The first part is the connection between adjacent nodes in each frame, denoted as $E_T = \left\{v_i^t v_j^t \mid (i, j) \in Q; t = 1, 2, ..., T\right\}$, where Q is the set of naturally connected joint pairs in the human body. The second part is the connection between the corresponding nodes of adjacent frames as $E_F = \left\{v_i^t v_i^{t+1} \mid i = 1, 2, ..., N; t = 1, 2, ..., T-1\right\}$. It should be noted here that the nodes are numbered to facilitate the construction of links between nodes, form the skeleton graph structure of the human body, and also build the connections between the corresponding nodes between adjacent frames. It is worth mentioning that the constructed skeleton graph structure is undirected and unordered.

In terms of the above-mentioned definition of the skeleton-based graph structure, a multi-layer GCN is constructed for extracting the spatial features of the skeleton structure. The adaptive global average pooling layer and Softmax classifier are then used to predict the action category based on the extracted features in this paper.

Based on [16], the spatial GCN calculation formula of the skeleton sequence at the tth frame can be expressed as in (1):

$$f_{out}\left(v_i^t\right) = \sum_{v_j \in B(v_i)} \frac{1}{Z_i^t\left(v_j^t\right)} f_{in}\left(v_j^t\right) \cdot w\left(l_i^t\left(v_j^t\right)\right) \tag{1}$$

where $f_{in}(\cdot)$ and $f_{out}(\cdot)$ denote the mapping rules of input and output, respectively. And v denotes the vertex of the graph. $B(v_i)$ is the sampling range, which is defined as the set of adjacent vertices of the target vertex v_i. $w(\cdot)$ denotes the weight function, which provides an initial weight vector for the input data, and there is a fixed number of the weight vectors. Z_i denotes the normalization term, which is equal to the cardinality of the corresponding subset. This item is added to balance the contribution of each subset to the output. $l_i(\cdot)$ is the mapping rule, which assigns a different weight vector to each different node.

The temporal domain convolution method can be directly utilized for extracting the temporal features of the skeleton data. However, it is complicated to implement graph convolution in the spatial dimension. To implement ST-GCN, (1) can be transformed into:

$$f_{out}(v_i) = \sum_{j=1}^{N} W f_{in}(v_j)\left(\Lambda^{-\frac{1}{2}} A_{ij} \Lambda^{-\frac{1}{2}} \circ M\right) \tag{2}$$

where N represents the total number of joint vertices in one frame. The definitions of f_{out} and f_{in} are similar to those in formula (1). A means an N-order adjacency matrix, where $A_{ij} = 1$ when vertexes v_i and v_j are adjacent in physical location; otherwise $A_{ij} = 0$. Λ is used to normalize A. Specifically, its elements can be expressed as $\Lambda_{ii} = \sum_{j=1}^{N} A_{ij} + \varepsilon$. In order to avoid invalid calculations, we refer to [6] and set ε to 0.001. Both W and M are learnable parameter matrices. And W is the weight vector of the 1×1 convolution operation with a size of $C_{in} \times C_{out} \times 1 \times 1$. C_{in} and C_{out} represent the numbers of channels of the input and output feature maps, respectively. M is used to adjust the importance level of each edge. And the operator $\circ$ denotes hadamard product.

The entire process of completing spatiotemporal graph convolution can be summarized as in formula (3):

$$f_{out}\left(v_i\right) = T_t\left[\alpha\left(\sum_{j=1}^{N} W f_{in}\left(v_j\right)\left(\Lambda^{-\frac{1}{2}} A_{ij} \Lambda^{-\frac{1}{2}} \circ M\right)\right)\right] \tag{3}$$

where $T_t\left[\cdot\right]$ is the temporal convolution layer, $\alpha\left(\cdot\right)$ denotes activation function.

3.2 Spatial adaptive residual graph convolution network

In this section, we will elaborate on each component of the spatial adaptive residual graph convolution network (SARGCN) in detail, and briefly explain the framework of the entire model.

The proposed basic network is mainly composed of two parts: spatial adaptive residual GCN and temporal convolutional network (TCN). The input skeleton sequence is composed of multiple frames, and the skeleton sequence of each frame is traditionally calculated based on a predefined graph when the graph convolution operation is performed. Perhaps this was not a good choice. The basic network framework we propose is shown in Fig. 1: Firstly, the skeleton data is preprocessed, and then the preprocessed skeleton data is input into K cascaded feature extraction modules to extract spatial and temporal features. Finally, the action category is output by a classifier. The specific structure of each feature extraction module is shown in Fig. 2. In the feature extraction module, there are two important blocks: SARGCN and TCN blocks, where the features in the spatial dimension of skeleton data are extracted through the SARGCN block, while the temporal features are obtained from the TCN block. First, the preprocessed skeleton data in the previous stage are fed to the SARGCN block for extracting the spatial features after downsampling. Next, through up-sampling, input to the TCN block to extract the temporal dimensional features. Finally, an attention block is used to improve the feature extraction ability of the model. The downsampling, through which the preprocessed data is entered into the SARGCN block, can reduce the number of model learning parameters, prevent over-fitting during training, and also expand the receptive field. In the temporal dimension, with the fixed number of neighbors of each node, the traditional three Conv2d-BN layers can be used to extract features, that is the same as ST-GCN. To stabilize the training, a residual connection is added to each SARGCN block. In the experiment, multiple feature extraction modules can be superimposed to achieve the best experimental results. Based on experience, K is set to 5, 6, and 7 in this paper, where the value of K also equals the number of GCN.

According to formula (2), the topological structure of the skeleton graph is actually determined by the adjacency matrix A and the mask matrix M, respectively. The adjacency matrix A indicates the connection relation between the nodes, and M indicates the strength of the connection between the nodes. In order to describe the adaptive graph structure, formula (2) can be rewritten as:

$$f_{out} = \sum_{j=1}^{N} W f_{in}\left(\Omega_{ij} + \Delta_{ij}\right) \tag{4}$$

where Ω_{ij} represents the hadamard product of the normalized adjacency matrix A and the mask matrix M. Δ_{ij} is added for the adaptive generation of the adjacency matrix, which will be described in detail next.

For graph structure data, Euclidean distance is no longer a good indicator of vertex similarity. And the distance metric needs to be adjusted adaptively along with the task and the features during training. In the article, Mahalanobis distance is introduced as an indicator of distance measurement. The

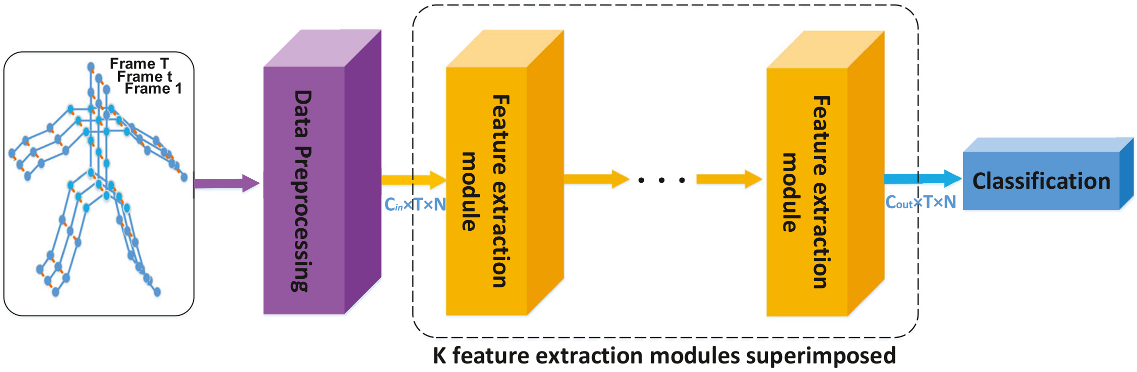
Frame T
Frame t
Frame 1
Data Preprocessing
C_{in}×T×N
Feature extraction module
. . .
Feature extraction module
C_{out}×T×N
Classification
K feature extraction modules superimposed

Fig. 1 Basic network framework

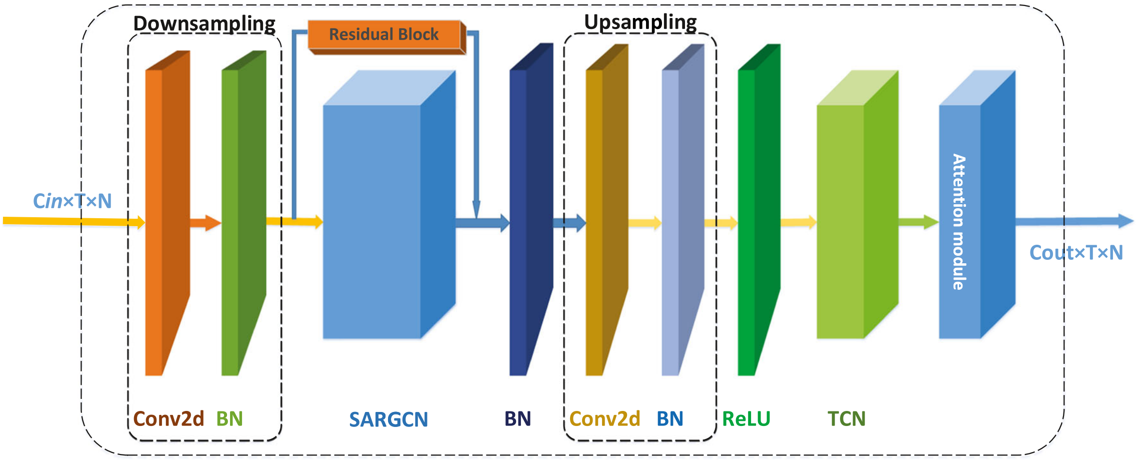

Fig. 2 Feature extraction module

Mahalanobis distance between nodes v_i and v_j can be calculated from the formula (5):

$$D\left(v_i, v_j\right) = \sqrt{\left(v_i - v_j\right)^T \Sigma \left(v_i - v_j\right)} \tag{5}$$

where $\Sigma = W^T W$, W is a trainable weight matrix in the SARGCN block, which is equivalent to the weighting function in formula (3), When $\Sigma = I$, $D\left(v_i, v_j\right)$ represents Euclidean distance. In order to determine whether there is a connection and evaluate the connection strength between two nodes, a Gaussian kernel is introduced :

$$G_{v_i,v_j} = e^{\left(-\frac{D^2\left(v_i,v_j\right)}{2\sigma^2}\right)} \tag{6}$$

where σ is a constant.

The Gaussian kernel represented by formula (6) is normalized to obtain (7):

$$\bar{G}_{v_i,v_j} = \frac{G_{v_i,v_j}}{\sum_{j=1}^{N} G_{v_i,v_j}} \tag{7}$$

The normalized value range of the matrix elements is between [0, 1] , and the normalized Gaussian function has a Softmax operation. Therefore, based on the above inference, Δ_{ij} can be expressed as in formula (8):

$$\Delta_{ij} = \bar{G}_{v_i,v_j} = Softmax\left(f_{in}^T W^T W f_{in}\right) \tag{8}$$

The overall structure of the SARGCN is shown in Fig. 3. Inspired by the good performance of Residual and AGCN models in the human action recognition based on skeleton data. Firstly, extract the skeleton features after downsampling in the spatial dimension. That is referred to as the Ω_{ij} operation and the Δ_{ij} operation in this article.

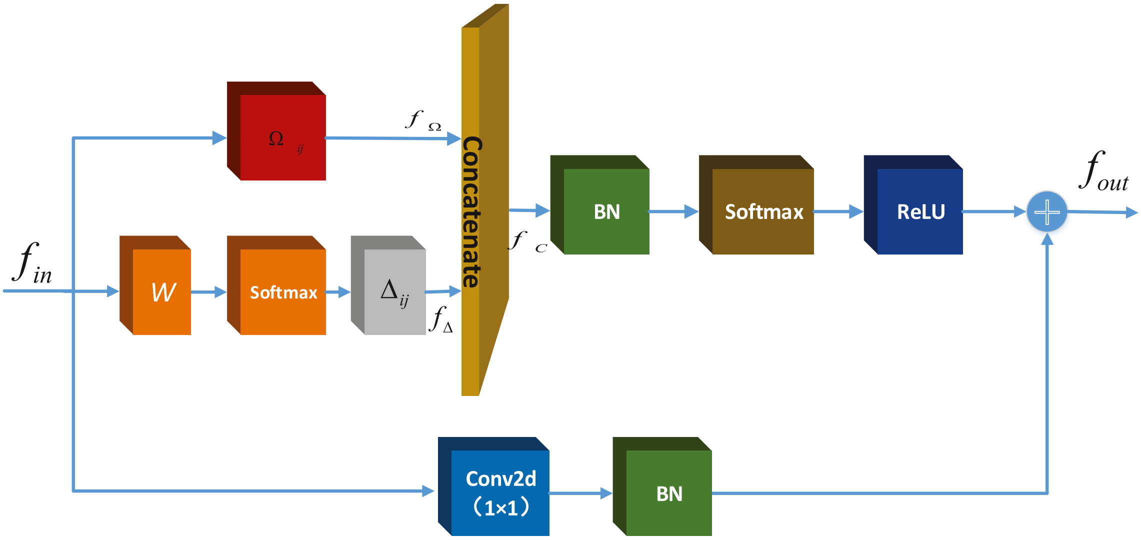

Fig. 3 SARGCN module with residual link

The two operations can be formulated as in (9) and (10), respectively.

$$f_{\Omega} = \sum_{j=1}^{N} Wf_{in} \cdot \Omega_{ij} \tag{9}$$

$$f_{\triangle} = \sum_{j=1}^{N} Wf_{in} \cdot Softmax\left(f_{in}^{T} W^{T} Wf_{in}\right) \tag{10}$$

Then, the features of f_{Ω} and $f_{\triangle}$ are concatenated and sequentially transmitted through a FC layer, a BatchNorm layer, a Softmax function, and a ReLu function. Finally, to improve the stability of the graph structure learned by the network, a residual connection is added to the entire module. Our proposed SARGCN can be formulated as:

$$f_c = Concat\left(f_{\Omega}, f_{\triangle}\right) \tag{11}$$

$$f_{out} = Res\left(ReLU\left(Softmax\left(BN\left(f_c\right)\right)\right), f_{in}\right) \tag{12}$$

According to formulas (11) and (12), it can be known that there is a key BatchNorm layer in the SARGCN module, through which the data are normalized after the feature concatenation. In this way, the traditional AGCN method is optimized so as to reduce the size of its model.

The cross-entropy loss is introduced to be the metric of the cost evaluation during training, and can be described as in formula (13):

$$L_{loss} = -\sum_{c=1}^{C} p_c log\left(\hat{p_c}\right) \tag{13}$$

where C is the number of behavior categories, p_c and $\hat{p_c}$ denote the one-hot vector of the ground truth, the prediction vector, respectively.

3.3 Attention module

Considering that human usually gives different attention perceptrons according to the important levels of things, an attention mechanism is also integrated into our network. Human perception usually selectively focuses on certain parts of the scene in order to obtain specific pieces of information. Skeleton data are a series of temporal sequences composed of a series of 3D coordinates that form an action. Different frames play different roles in importance levels in the process of action recognition. For instance, in the action of brushing teeth, the action is similar in most frames but changes greatly in only a few frames, which are also the key to identifying the action. Inspired by this, we introduce an attention mechanism to enhance the weights of frames carrying key information, thereby improving the recognition accuracy of the model.

In this paper, an attention mechanism is introduced following the feature extraction module. On the basis of the attention mechanism [3, 28, 29], as shown in Fig. 4, we design a different and adaptive weight matrix to each frame according to the different importance levels of each frame in the entire action sequence from both spatial and temporal domains. In the attention module, The feature information is first fed into the fully connected network with an adaptive average pooling layer. Next, the attention matrix is calculated by a fully connected layer with the BatchNorm layer and the ReLU function. Finally, a Softmax layer is used to determine the key actions in the key action frames. More specifically, the implementation of the attention module can be expressed as an equation (14):

$$f_{att\_out} = Softmax\left(ReLU\left(X \circ FCN\left(f_{att\_in}\right)\right)\right) \tag{14}$$

where f_{att_in} and f_{att_out} denote the input and output of the attention module, respectively. X represents the attention parameter matrix which is adjustable during learning, and has the same size as the output in formula (3). And $\circ$ denotes hadamard product.

4 Experiments

In this section, two public large-scale action recognition datasets are used to evaluate our model and a comparison is made between our method and the state-of-the-art methods. Moreover, multiple sets of experiments are conducted for evaluating the impact of the number K of layers in the network on the performance of the model.

4.1 Datasets

NTU RGB+D 60 [18] is a large-scale dataset that is captured indoors and widely used in the recognition of skeleton-based actions. The dataset is composed of more than 56000 human action videos recorded by three Kinect cameras, including 60 actions. Some interactions between two subjects are depicted in the last 10 classes of the actions. These human actions are performed by 40 volunteers. There exist at most two volunteers in each frame of videos, and each skeleton consists of 25 joints denoted by three-dimensional coordinates. We evaluate the model based on two benchmark sub-datasets: cross-subject (X-Sub) and cross-view (X-View). 1)X-Sub: this sub-dataset is performed by two groups of 20 volunteers each. These two groups of volunteers complete a total of 56880 videos, where 40320 samples are used for training and 16560 samples for testing. 2)X-View: this sub-dataset is taken by three cameras, where 37920 videos from cameras 2 and 3 are used as training samples, and 18960 videos from camera 1 are used as testing samples.

NTU RGB+D 60 dataset is further extended to be NTU RGB+D 120 dataset [19] which is so far the largest action

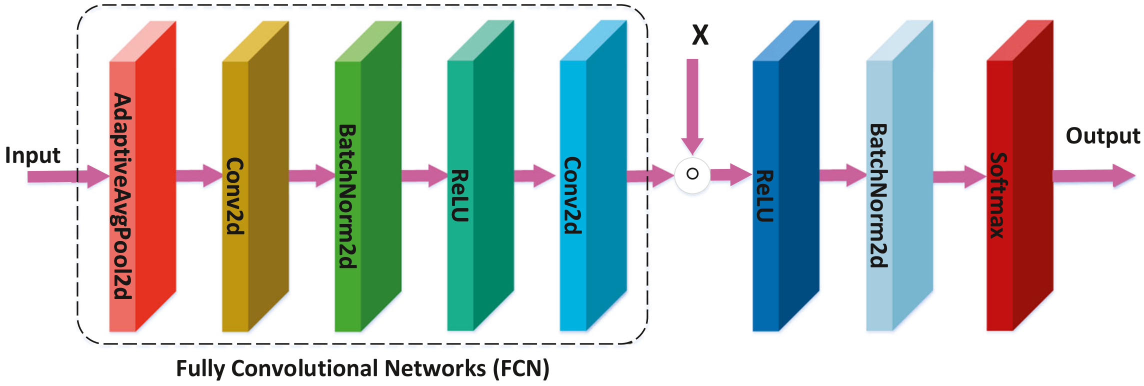

Fig. 4 Attention module

recognition dataset captured indoors. This extended dataset includes 114480 videos performed by 106 volunteers from 155 viewpoints and 32 setup IDs, generating 120 actions. 532 bad samples of this dataset should be ignored in all experiments. Similarly, cross-subject (X-Sub120) and cross-setup (X-Set120) are recommended to be the two benchmark evaluations. 1) X-Sub120: In the total of 114480 videos performed by 106 subjects, one group of 53 subjects completed 63026 videos for training, and the other group of 53 subjects completed 50922 videos for testing, respectively. 2) X-Set120: the 32 IDs are divided into two groups with 16 IDs each, of which 16 ID sequences form the training set, and the two groups of ID sequences form the 54471 samples for training and the 59477 samples for testing, respectively.

4.2 Experiment settings

All our experiments are conducted in the Pytorch framework [30] and on NVIDIA GTX 1080Ti GPU. The stochastic gradient descent (SGD) rule with the Nesterov momentum of 0.9 and the weight decay of 0.0002 is adopted to optimize the network. In the spatial dimension, the size of the convolution kernel is 9 × 1. In the temporal dimension, the size of the convolution kernel is 1 × 1. The number of frames of each skeleton sequence is set to be 300, and all 0s are filled at the end frames for video samples with less than 300 frames. The batch size and the maximum iterating epoch are set to be 16 and 70, respectively. The learning rate is initially set to 0.01, which is increased by 0.01 each epoch during the first 10 epochs, and remains 0.1 during the second 10 epochs. Then by referring to [25], the learning rate decays by a factor of 10 from the 21st epoch to the 50th epoch. From the 51st epoch to the 70th epoch, the learning rate decays again by a factor of 10.

4.3 Comparison

To evaluate the performance of SARGCN, we make a comparison between our proposed SARGCN and other state-of-the-art approaches of skeleton-based action recognition on the NTU RGB+D dataset. The methods for comparison can be divided into two categories: traditional methods (such as RNN-based methods and CNN-based methods) and GCN-based methods.

The results are shown in Tables 1 and 2. Our SARGCN benefits from the robustness of AGCN to extract global spatial information and the optimization of traditional

Table 1 Accuracy comparison between our proposed six-layer SARGCN and the traditional methods on NTU RGB+D 60 and NTU RGB+D 120

Method	Param.	X-Sub	X-View	X-Sub120	X-Set120
Hierarchical RNN [5]	-	59.10%	64.00%	-	-
Dynamic skeletons [31]	-	60.23%	65.22%	-	-
ST-LSTM+Trust Gate [32]	-	69.20%	77.70%	55.00%	57.90%
Two-stream RNNs [33]	-	71.30%	79.50%	-	-
STA-LSTM [34]	-	73.40 %	81.20%	-	-
Two-Stream3DCNN [35]	-	66.80%	72.60%	-	-
Clips+CNN+MTN [13]	-	79.60%	84.80%	-	-
3scale ResNet152 [15]	-	85.00%	92.30%	-	-
Proposed Method	1.09M	88.91%	94.83%	83.81%	85.11%

Table 2 Accuracy comparison among our proposed six-layer SARGCN and the other state-of-the-art GCN methods on NTU RGB+D 60 and NTU RGB+D 120

Method	Param.	FLOPs	X-Sub	X-View	X-Sub120	X-Set120
ST-GCN [16]	3.10M	16.32G	81.50%	88.30%	70.70%	73.20%
RA-GCN [36]	6.21M	32.80G	85.90%	93.50%	74.60%	75.30%
AS-GCN [24]	6.99M	26.76G	86.80%	95.10%	82.50%	84.20%
2s-AGCN [6]	9.94M	37.32G	88.50%	95.10%	82.50 %	84.20%
PL-GCN [37]	20.70M	-	89.20%	90.50%	-	-
PA-ResGCN-B19 [26]	3.64M	18.52G	90.90%	96.00%	87.30%	88.30%
ST-TR [38]	-	-	89.90%	96.10%	81.90%	84.10%
SAGN [39]	1.83M	-	89.20%	94.20%	82.10%	83.80%
2s-PST-GCN [40]	1.84M	-	88.68 %	95.10%	-	-
FGCN [41]	-	-	90.20 %	96.30%	85.40%	87.40%
DD-GCN [42]	-	-	88.90%	95.80%	84.90%	86.00%
4s-HybridNet [43]	-	-	91.40%	96.90%	87.50%	89.00 %
Proposed Method	1.09M	5.37G	88.91%	94.83%	83.81%	85.11 %

AGCN methods, and its results outperform those of traditional methods significantly.

According to Tables 1 and 2, the following results can be clearly found:

(1) On the NTU RGB+D 60 (120) dataset divided into X-Sub (X-Sub120) and X-View (X-Set120), the SARGCN recognition accuracies reach up to 88.91% (83.81%) and 94.83% (85.11%), respectively.

(2) Compared with the ST-GCN [16], the recognition accuracies of our SARGCN have been improved by 7.41% (13.11%) and 6.53% (11.91%) under X-Sub (X-Sub120) and X-View (X-Set120), respectively.

(3) To our best knowledge, compared with the highest accuracy of the current main-stream methods, although the accuracy of SARGCN recognition is not the best, the number of model parameters is currently the least.

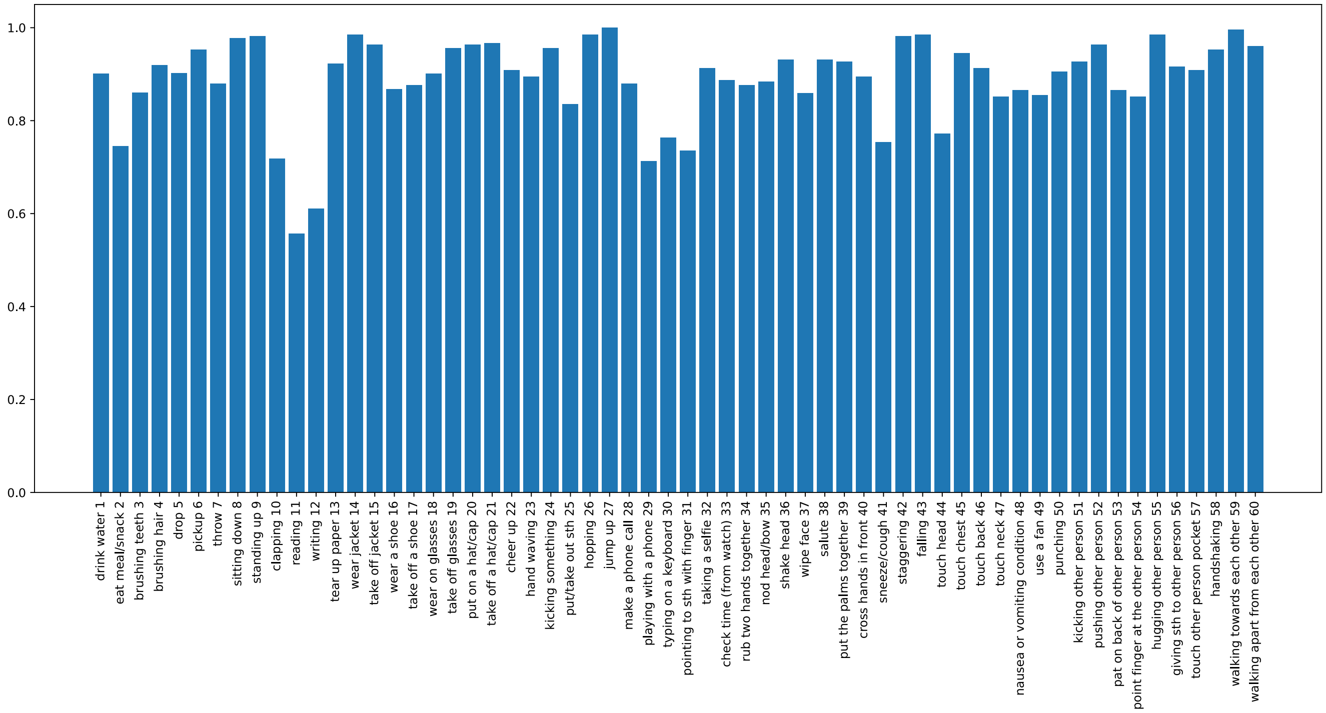

Fig. 5 The accuracy comparison result of each category on the NTU-RGB+D 60 X-Sub datasets. The horizontal and vertical axes denote the category and the accuracy, respectively

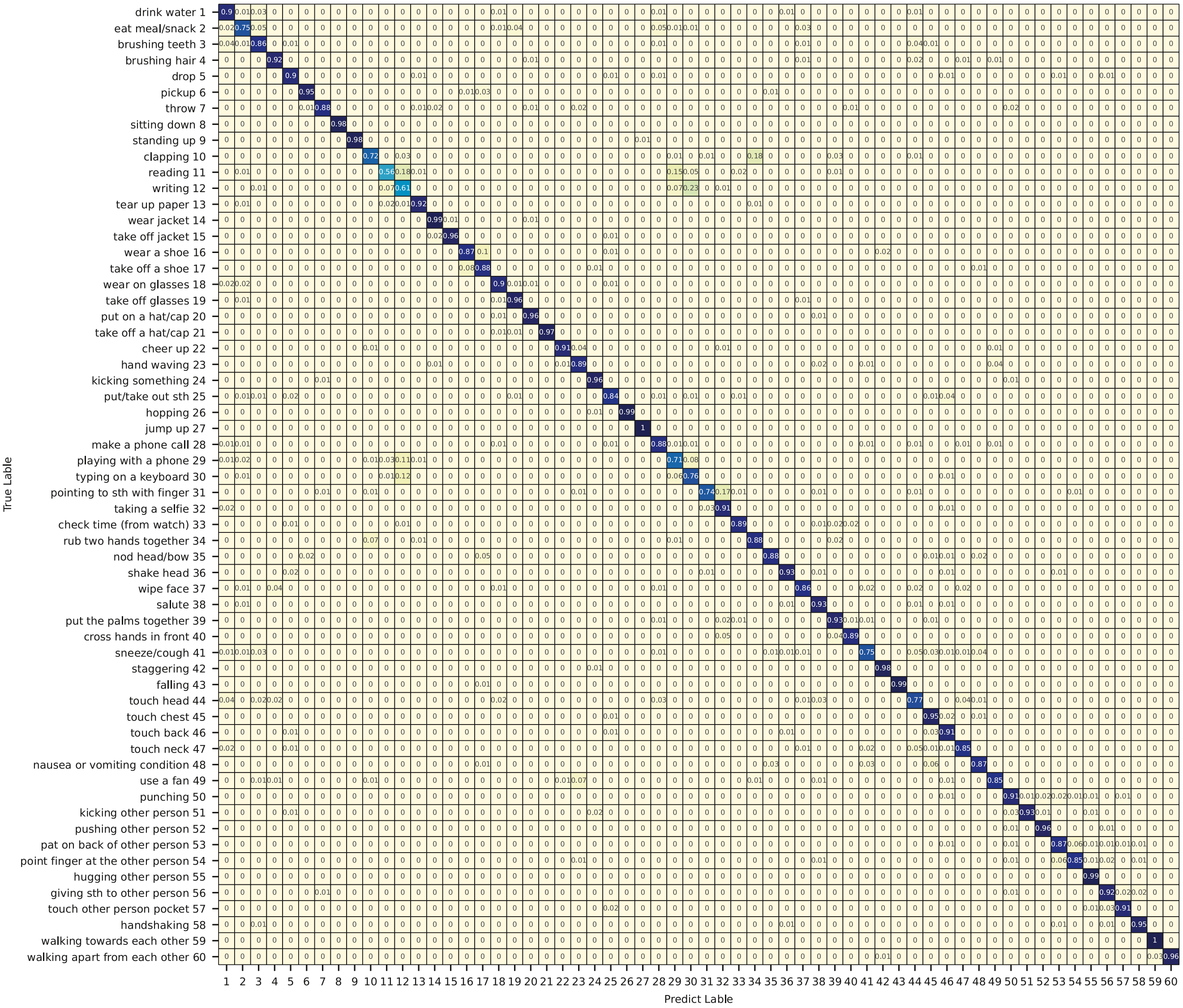

Fig. 6 Confusion matrix on the NTU-RGB+D 60 X-Sub datasets

(4) Compared with the PA-ResGCN-B19 [26] method, the parameter amount of the SARGCN model is only one-third of that.

Tables 1 and 2 list the performance comparison between the proposed method and other state-of-the-art non-GCN and GCN methods on the NTU dataset, respectively. It shows that our proposed model achieves an excellent performance of 88.91% (83.81%) and 94.83% (85.11%) on NTU RGB+D 60 (120). There are three experiments in each group, and the standard error of each group is no more than 0.15. Furthermore, we calculate the recognition accuracy and confusion matrix of our network on X-Sub dataset, as shown in Figs. 5 and 6.

In Table 2, compared with [38] and [39], although our accuracy under the X-Sub setting is slightly lower than theirs, the accuracy under the more complicated X-Sub120 setting is 1.91% and 1.71% higher than theirs, respectively. Meanwhile, our method also outperforms theirs by 1.01% and 1.31% in accuracy for the X-Set. Compared with [40], the accuracy is improved by 0.23% on the X-Sub setting, and the number of parameters of the model is reduced by 0.75M. Although the experimental results of PA-ResGCN-B19 [26], FGCN [41], DD-GCN [42] and HybridNet [43] performed slightly better than ours, the structure of our model is simpler than theirs, and our model has fewer parameters. For example, compared with [26], the number of parameters in our model is reduced by 2.55M. Besides,

Table 3 Comparison of different layers of networks (without attention module)

Method	Param.	FLOPs	X-Sub	X-View	X-Sub120	X-Set120
5 layers network	0.57M	4.69G	88.11%	93.41%	82.56%	83.00%
6 layers network	0.71M	5.34G	88.20%	94.32%	83.16%	83.61%
7 layers network	0.77M	5.91G	88.07%	94.03%	82.81%	84.05%

Table 4 Comparison of different layers of networks (without residual connection)

Method	Param.	FLOPs	X-Sub	X-View	X-Sub120	X-Set120
5 layers network	0.69M	3.89G	85.78%	91.75%	78.72%	79.91 %
6 layers network	0.96M	4.55G	85.01%	91.13%	76.88%	78.09%
7 layers network	1.04M	5.12G	82.90%	90.44%	75.76%	76.88%

Table 5 Comparison of different layers of networks (with attention module and residual connection)

Method	Param.	FLOPs	X-Sub	X-View	X-Sub120	X-Set120
5 layers network	0.82M	4.71G	88.75%	94.59%	83.52%	84.90%
6 layers network	1.09M	5.37G	88.91%	94.83%	83.81%	85.11%
7 layers network	1.17M	5.94G	87.74 %	93.38 %	80.57 %	84.06%

Table 6 Comparison of cross-validation results under X-View benchmark

Method	Param.	FLOPs	without cross-validation	cross-validation
5 layers network	0.82M	4.71G	94.59 %	91.57 %
6 layers network	1.09M	5.37G	94.83 %	92.28 %
7 layers network	1.17M	5.94G	93.38 %	92.42 %

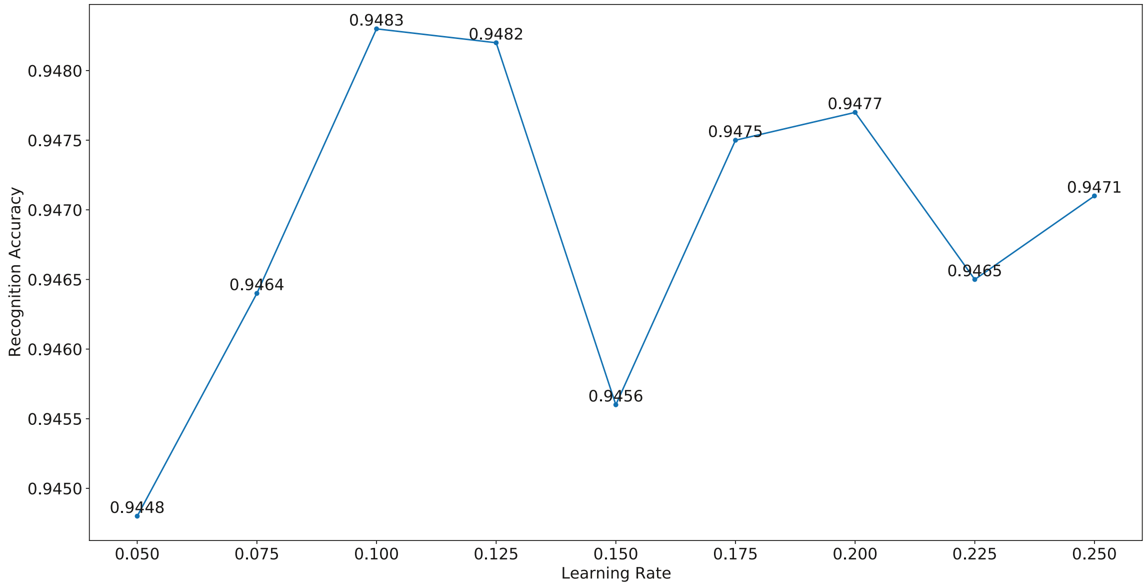

Fig. 7 Recognition accuracy under different learning rates

the addition of some extra modules to their model, increased the complexity of the model and thus affecting the inference speed of the model. Our model has a faster inference speed with 5.37G FLOPs. The main reason why our model is slightly inferior to [26] in recognition accuracy is that the processed skeleton graph with 25 nodes and their connecting edges are input into the network as a whole, and the adjacency matrix is generated adaptively as a whole in our method. This will inevitably ignore some subtle movements, thereby affecting the recognition results. As shown in Fig. 5, it can be clearly found that both recognition accuracies of actions in reading and writing are much lower than those of other categories, because the variation ranges of these two categories of actions are very slight, and it is not easy for the general model to find changes in subtle actions. In view of the accuracy and the number of parameters, the proposed method achieves the best result.

4.4 Ablation study

In this section, we further investigate the impact of the model scale on recognition accuracy. In the experiment, we keep the other parameters unchanged and only change the number of the network layers K. On the basis of a large number of our previous experiments, we select three models with 5, 6, and 7 layers separately for comparison in the experiment. Lots of experiments demonstrate that the best accuracy can be achieved when the number of network layers is 6. The evaluation indicators are shown individually in Tables 3, 4 and 5.

Tables 3 and 4 show the results without the attention module and without residual connection for each SARGCN, respectively. To validate the contribution of the added attention module and residual connections, repeat the above experiment after adding the attention module and residual connections. The experimental results is shown in Table 5. Comparing the results of the three experiments, the recognition accuracy has been improved after adding the attention module and residual connections. Similarly, the recognition effect is also the best under the six-layer network structure in each of the three experiments.

In order to further evaluate the validity of our model, we perform cross-validation on the dataset of X-View benchmark, and the validation results are shown in Table 6. To be specific, we try re-partitioning the dataset into 5 parts, and adopt 5-fold cross validation in our further experiments. As far as we know, almost all the existing works followed the original split method of the authors of the dataset. Therefore, our experimental result with cross-validation is only compared with that of our own methods rather than those of other SOTA methods. Meanwhile, we set different learning rates in the X-View benchmark for verification, as shown in Fig. 7, where the experimental accuracy reaches the best when the learning rate is equal to 0.1.

5 Conclusion

A novel adaptive spatial residual graph convolutional network (SARGCN) is proposed for action recognition based on the skeleton information. In our model, without the constraint of fixed topological structure, the feature extraction performance and generalization ability could be largely enhanced by a learnable parameter matrix. In the meantime, the residual connection is introduced into the model so that model degradation could be eliminated and the computational complexity of the model could be reduced. At last, the attention module is added to promote the model's extraction of spatial features and achieve a significant effect. In terms of the number of parameters, to our best knowledge, our model has achieved the most effective result so far.

Of course, there is still a certain gap in recognition accuracy for our model in comparison to the state-of-the-art action recognition model based on skeleton data. The reason is that we only take the overall information of the skeleton structure in the spatial dimension in our approach. In our future work, we will further investigate the feature extraction of each part of the skeleton, and consider how to re-establish a spatiotemporal graph model on the basis of the comprehensive analysis of the spatiotemporal feature information in the spatial and temporal dimensions.

Author Contributions Conceptualization, methodology, and writing original draft, Qilin Zhu; writing—review, editing, and supervision, Hongmin Deng. All authors have read and agreed to the published version of the manuscript.

Funding This work was partially supported by the National Natural Science Foundation of China under Grant 62020106010.

Declarations

Conflict of Interests The authors declare no conflict of interest.

References

1. Niebles JC, Wang H, Li FF (2008) Unsupervised learning of human action categories using spatial-temporal words. Int J Comput Vis 79:299–318
2. Niebles JC, Li FF (2007) A hierarchical model of shape and appearance for human action classification, 2007 IEEE Conference on computer vision and pattern recognition. 1–8
3. Song S, Lan C, Xing J, Zeng W, Liu J (2017) An end-to-end spatio-temporal attention model for human action recognition from skeleton data. AAAI Conference on artificial intelligence, north america. 4263–4270
4. Vemulapalli R, Arrate F, Chellappa R (2014) Human action recognition by representing 3D skeletons as points in a lie group, 2014 IEEE conference on computer vision and pattern recognition. 588–595
5. Du Y, Wang W, Wang L (2015) Hierarchical recurrent neural network for skeleton based action recognition, 2015 IEEE Conference on computer vision and pattern recognition (CVPR). 1110–1118
6. Shi L, Zhang Y, Cheng J, Lu H (2019) Two-stream adaptive graph convolutional networks for skeleton-based action recognition. Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 12026–12035
7. Li S, Li W, Cook C, Zhu C, Gao Y. (2018) Independently recurrent neural network (IndRNN): building a longer and deeper RNN. 2018 IEEE conference on computer vision and pattern recognition. 5457–5466
8. Zhang P, Lan C, Xing J, Zeng W, Xue J, Zheng N (2017) View adaptive recurrent neural networks for high performance human action recognition from skeleton data, 2017 IEEE international conference on computer vision (ICCV). 2136–2145
9. Si C, Jing Y, Wang W, Wang L, Tan T (2018) Skeleton-based action recognition with spatial reasoning and temporal stack learning. Proceedings of the european conference on computer vision (ECCV), 103–118
10. Zhu W, Lan C, Xing J, Zeng W, Li Y, Shen L, Xie X (2016) Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks, Proceedings of the AAAI conference on artificial intelligence 30(1)
11. Kim TS, Reiter A (2017) Interpretable 3d human action analysis with temporal convolutional networks, 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 1623–1631
12. Li C, Hou Y, Wang P, Li W (2018) Multiview-based 3-D action recognition using deep networks. IEEE Transactions on Human-Machine Systems 49(1):95–104
13. Ke Q, Bennamoun M, An S, Sohel F, Boussaid F (2017) A new representation of skeleton sequences for 3d action recognition. Proc IEEE conf comput vis pattern recognit, 3288–3297
14. Cao C, Lan C, Zhang Y, Zeng W, Lu H, Zhang Y (2018) Skeleton-based action recognition with gated convolutional neural networks. IEEE Trans Circuits Syst Video Technol 29(11):3247–3257
15. Li B, Dai Y, Cheng X, Chen H, Lin Y, He M (2017) Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN, 2017 IEEE international conference on multimedia & expo workshops (ICMEW). IEEE, 601–604
16. Yan S, Xiong Y, Lin D (2018) Spatial temporal graph convolutional networks for skeleton-based action recognition, 32nd AAAI Conference on artificial intelligence, New Orleans. LA. 02-07, 7444–7452
17. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. Proc IEEE conf comput vis pattern recognit, 770–778
18. Shahroudy A, Liu J, Ng TT, Wang G, NTU RGB+ D (2016) A large scale dataset for 3d human activity analysis. Proc IEEE Conf comput vis pattern recognit, 1010–1019
19. Liu J, Shahroudy A, Perez M, Wang G, Duan LY, Kot AC, NTU RGB+ D (2019) 120: A Large-scale benchmark for 3d human activity understanding[J]. IEEE Trans Pattern Anal Mach Intell 42(10):2684–2701
20. Kipf TN, Welling M (2016) Semi-supervised classification with graph convolutional networks. arXiv:1609.02907
21. Duvenaud D, Maclaurin D, Iparraguirre JA, Bombarelli RG, Hirzel T, Guzik AA, Adams RP (2015) Convolutional networks on graphs for learning molecular fingerprints. In: Proceedings of the 28th international conference on neural information processing systems - Volume 2 (NIPS'15), MIT Press, Cambridge, MA, USA, pp 2224–2232
22. Atwood J, Pal S, Towsley D, Swami A (2017) Sparse diffusion-convolutional neural networks. arXiv:1710.09813
23. Hamilton WL, Ying R, Leskovec J (2017) Inductive representation learning on large graphs. Proceedings of the 31st international conference on neural information processing systems, 1025–1035
24. Li M, Chen S, Chen X, Zhang Y, Wang Y, Tian Q (2019) Actional-structural graph convolutional networks for skeleton-based action recognition. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 3595–3603
25. Song YF, Zhang Z, Shan C, Wang L (2020) Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition. Proceedings of the 28th ACM International Conference on Multimedia, 1625–1633
26. Shi L, Zhang Y, Cheng J, Lu H (2020) Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. In: IEEE Transactions on Image Processing, vol 29, pp 9532–9545
27. Zhang X, Xu C, Tian X, Tao D (2019) Graph edge convolutional neural networks for skeleton-based action recognition. In: IEEE Transactions on Neural Networks and Learning Systems, vol 31, pp 3047–3060
28. Baradel F, Wolf C, Mille J (2017) Human action recognition: Pose-based attention draws focus to hands. Proceedings of the IEEE International Conference on Computer Vision Workshops, 604–613
29. Si C, Chen W, Wang W, Wang L, Tan T (2019) An attention enhanced graph convolutional lstm network for skeleton-based action recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1227–1236
30. Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin Z, Desmaison A, Antiga L, Lerer A (2017) Automatic differentiation in pytorch. In: Proc. Adv. Neural Inf Proc.ss. Syst. Workshops, pp 1–4
31. Bar EO, Trivedi MM (2013) Joint angles similarities and HOG2 for action recognition. Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 465–470
32. Liu J., Shahroudy A, Xu D, Wang G (2016) Spatio-temporal LSTM with trust gates for 3d human action recognition, european conference on computer vision. Springer, Cham, pp 816–833
33. Ji Y, Cheng H, Zheng Y, Li H (2015) Learning contrastive feature distribution model for interaction recognition. J Vis Commun Image Represent 33:340–349
34. Liu B, Ju Z, Liu H (2018) A structured multi-feature representation for recognizing human action and interaction. Neurocomputing 318:287–296

35. Liu H, Tu J, Liu M (2017) Two-stream 3d convolutional neural network for skeleton-based action recognition. arXiv:1705.08106
36. Song Y, Zhang Z, Wang L (2019) Richly activated graph convolutional network for action recognition with incomplete skeletons, 2019 IEEE International Conference on Image Processing (ICIP), 1–5
37. Huang L, Huang Y, Ouyang W, Wang L (2020) Part-level graph convolutional network for skeleton-based action recognition. Proceedings of the AAAI conference on artificial intelligence 34(07):11045–11052
38. Plizzari C, Cannici M, Matteucci M (2021) Spatial temporal transformer network for skeleton-based action recognition, pattern recognition. ICPR International workshops and challenges, 694–701
39. Fu Z, Liu F, Zhang J, Wang H, Yang C, Xu Q, Qi J, Fu X, Zhou A (2021) SAGN: Semantic Adaptive graph network for Skeleton-Based human action recognition. In: Proceedings of the 2021 international conference on multimedia retrieval (ICMR'21), pp 110–117
40. Heidari N, Iosifidis A (2021) Progressive Spatio-Temporal graph convolutional network for skeleton-based human action recognition, ICASSP 2021 - 2021 IEEE international conference on acoustics, speech and signal processing (ICASSP), 3220–3224
41. Yang H, Yan D, Zhang L, Sun Y, Li D, Maybank SJ (2022) Feedback graph convolutional network for skeleton-based action recognition. IEEE Trans Image Process, 164–175
42. Chen S, Xu K, Mi ZJ, Jiang XH, Sun TF (2022) Dual-domain graph convolutional networks for skeleton-based action recognition. Machine Learning
43. Yang WJ, Zhang JL, Cai JJ, Xu ZY (2022) Hybridnet: Integrating GCN and CNN for skeleton-based action recognition. Applied Intelligence