Multimedia Tools and Applications (2024) 83:73289–73304
https://doi.org/10.1007/s11042-023-17176-x

1232: HUMAN-CENTRIC MULTIMEDIA ANALYSIS

Enhanced decoupling graph convolution network for skeleton-based action recognition

Yue Gu[1,2] · Qiang Yu[1] · Wanli Xue[1,2]

Received: 10 January 2023 / Revised: 14 September 2023 / Accepted: 19 September 2023 /
Published online: 26 October 2023

Abstract

In skeleton-based action recognition, graph convolution networks have been widely applied and very successful. However, because graph convolution is a local operation with a small field of perception, it cannot investigate well for the connections between joints that are far apart in the skeleton graph. In addition, graph convolution makes all channels share the same adjacency matrix, which causes the topology learned to be the same among different channels, which limits the ability of graph convolution to learn topological information. In this paper, we propose an enhanced decoupling graph convolution network that effectively expands the perceptual field of the graph convolution by adding additional graphs, and the decoupled feature fusion mechanism increases its expressive power. In addition, we introduce an attention mechanism in the model to obtain the important elements in the whole feature map from both spatial and temporal dimensions simultaneously, so that the graph convolution can focus on the important elements more precisely and efficiently and suppress the influence of irrelevant elements on the model performance. To validate the effectiveness and advancedness of the proposed model, we conducted extensive experiments on three large datasets: NTU RGB+D 60, NTU RGB+D120 and Northwestern-UCLA. On the NTU RGB+D 60 dataset, the accuracy of our model archieves 91.6% and 96.5% on the two protocols.

Keywords Action recognition · Graph convolution networks · Attention mechanism

✉ Yue Gu
guyue@email.tjut.edu.cn

✉ Wanli Xue
xuewanli@email.tjut.edu.cn

Qiang Yu
yu@stud.tjut.edu.cn

[1] Key Laboratory of Computer Vision and System (Ministry of Education), School of Computer Science and Engineering, Tianjin University of Technology, Tianjin 300384, China

[2] Engineering Research Center of Learning-Based Intelligent System (Ministry of Education), Tianjin University of Technology, Tianjin 300384, China

1 Introduction

Human action recognition are used in a variety of fields, including intelligent video surveillance, athlete-assisted training, and virtual reality. Both the skeleton-based and the RGB-based methods are capable of recognizing human body actions. A lot of study and focus has been put on action recognition techniques based on skeleton data in particular, because they are robust to changing environments and intricate backgrounds. Additionally, skeleton data is more compact than RGB data since it just includes the spatial coordinates of the human body. Skeleton data are compiled into a a sequence of coordinate vectors [1–5] or pseudo-image [6–9] for feature extraction in early deep-learning-based action recognition methods. However, as skeleton data are topological data and these methods neglect the inherent association of topological data, their accuracy has been quite subpar. Yan et al. [10] first applied a spatial-temporal graph convolution network (ST-GCN) to skeleton data, in order to extract the associations within human joints and between various frames in both the temporal and spatial dimensions. It enabled graph convolution to obtain the intrinsic dependencies between human joints.

However, ST-GCN has some shortcomings: (1) The predefined adjacency matrix in ST-GCN only considers the connection relationship between adjacent nodes, while many actions of the human body not only produce connections in adjacent joints, but sometimes joints that are far apart also have strong connections. For example, the joints of the hand and head in the headphone wearing action and the joints of the hand and foot in the shoe removal action have strong connections between them although they are far apart in the human body structure, and it is difficult to have a comprehensive understanding of these actions if the connections between the two parts are not effectively obtained. Although ST-GCN overlays multiple graph convolution layers to capture these connections, because the graph convolution is a local operation, the connections between distant joints can only be acquired indirectly through many intermediate joints, which hinders information exchange, brings redundant computation, is very inefficient in computation, still does not effectively expand its perceptual field, and increases the difficulty of network optimization. In addition, the graph convolution method proposed in ST-GCN makes all channels share the same adjacency matrix, and the topology learned by the graph convolution is the same among different channels, which limits the ability of the graph convolution to learn topological information as well as the expression ability. (2) Since attention mechanisms have been shown to be effective in many computer vision tasks, especially for the study of skeleton-based action recognition. In the spatial structure of the skeleton, different joints of the human body occupy different importance in different actions. In the temporal dimension, each frame occupies different importance in different actions because each phase of the action has different importance to the whole action. The attention mechanism can help the neural network to better capture this importance information. However, no attention mechanism is added in ST-GCN, which leads to ST-GCN cannot focus on the important parts of different frames and different joints effectively.

In the context of human action recognition, spatial and temporal information are often intertwined, and capturing the connections between distant joint points also becomes crucial. To enhance the ability to aggregate spatial features and extract relevant spatio-temporal attention information, we introduce two modules: the enhanced decoupling graph convolution (ED-GC) module and the spatial-temporal union attention (STA) module. These modules, along with the temporal convolution module, form the enhanced decoupling graph convolution network (ED-GCN), enabling improved representation for accurate human action recognition. ED-GCN has two main contributions: (1) The ED-GC module add two additional graphs on basis of the adjacency graph, expanding the perceptual field of graph convolution. It also decouples the graph convolution operations and groups the channels so that each group of channels has an independent trainable adjacency matrix to extract information, effectively increasing the expressiveness and flexibility of the graph convolution and improving its ability to learn topological information. (2) The STA module captures important information from the input features in spatial and temporal elements within different channels, highlighting important features with high contributions in space and time, enabling the model to focus more effectively on these important features and suppressing the influence of useless features.

Our extensive experiments on three large datasets: NTU RGB+D 60 [11], NTU RGB+D120 [12] and Northwestern-UCLA [13] demonstrate that our model achieved superior performances.

2 Related work

2.1 Graph convolution network

Convolutional neural networks can successfully process Euclidean data, such as images, but it cannot directly handle non-Euclidean data, such as skeleton graphs. In recent years, skeleton data processing has been guided by graph convolution networks (GCNs) [14–22], a method which is highly good at handling non-Euclidean data. The two main kinds of GCNs are spectral perspective [14, 20, 21, 21, 22] and spatial perspective methods [17–20]. (1) Spectral perspective methods is to apply the Laplace matrix to the graph and study the properties of the graph according to its eigenvalues and eigenvectors. Considering the nodes' own properties of the graph as signals on the graph, the convolution can be defined as the multiplication of the signal and the filter in the Fourier domain. However, it is computationally costly. (2) Spatial perspective methods is based on the spatial connection between nodes, using the aggregation information of neighboring nodes. By defining an aggregation function to aggregate the central node and its neighboring nodes. Here we use the spatial perspective method to implement a graph convolution neural network.

2.2 Skeleton-based action recognition

Skeleton-based action recognition has garnered significant attention due to the rich information conveyed by human body dynamics, enabling the understanding of intricate human behaviors. Early approaches to modeling skeletons relied on hand-crafted components or traversal rules, which resulted in limited expressiveness and generalization difficulties, gradually fade out from the stage of frontier research.

Yan et al. [10] introduce ST-GCN, a model for dynamic skeletons. By automatically learning spatial and temporal patterns from data, ST-GCN overcame previous constraints, leading to enhanced expressive power and stronger generalization capability. Traditional GCN-based approaches utilized manually set and fixed graph topologies, limiting adaptability. To address this, Shi et al. [23] introduce a two-stream adaptive GCN (2s-AGCN). By learning graph topology and incorporating second-order information of skeletons, the recognition accuracy of their method is significantly improved. To further heighten the performance, Shi et al. [24] pioneer the multi-stream attention-enhanced adaptive graph convolutional neural network (MS-AAGCN) dedicated to skeleton-based action recognition. They also offer a distinctive representation by converting the skeleton data into a directed acyclic graph, grounded in the kinematic relationships existing between joints and bones within the human body [25]. In a quest to delve deeper into action-specific latent dependencies, Li et al. [26] propose an actional-structural graph convolution network (AS-GCN) with the A-link inference module, an encoder-decoder structure engineered to directly capture actional links from actions. Liu et al. [27] introduce a MS-G3D feature extractor to enhance long-range modeling and facilitate direct information propagation within the spatial-temporal graph. To enhance the model to leverage temporal dependencies between non-continuous frames and varying sequence lengths. Zhang et al. [28] introduce a Spatial Attentive and Temporal Dilated Graph Convolutional Network (SATD-GCN), which utilizes self-attention to select relevant body joints, and extracts temporal features across time scales. Chi et al. [29] focus on embedding skeleton information into latent representations, proposing InfoGCN with an information bottleneck-based learning objective and attention-based graph convolution. Duan et al. [30] introduce PoseConv3D, a 3D heatmap-based approach that addressed robustness, scalability, and generalization challenges. While recent trends leaned toward deep feedforward neural networks for joint coordinate modeling, considerations of computational efficiency were lacking. To bridge this gap, Zhang et al. [31] proposed a semantics-guided neural network (SGN) for skeleton-based action recognition, enhancing feature representation by explicitly incorporating high-level joint semantics. Furthermore, Cheng et al. [32] tackled the limitations of heavy computational complexity in GCN-based methods by introducing a novel shift graph convolutional network (Shift-GCN). This approach achieved superior performance with significantly reduced computational requirements. While these methods exhibit promising performance, there remains a need for increased focus on the interplay between spatial and temporal information, as well as the linkage between distant joint points. Additionally, the equilibrium between recognition performance and computational efficiency also require further attention.

3 Method

3.1 Preliminaries

3.1.1 Notations

The human skeleton is represented as a graph, where the joints are the vertices and the bones are the edges. The graph is represented as $G = (V, E)$, where $V = \{v_1, v_2, ..., v_N\}$ denotes the set of N joints and E denotes the set of edges of the skeleton represented by the adjacency matrix $A \in \mathbb{R}^{N \times N}$. If there is a skeleton connection from joint v_i to v_j, then $A_{i,j} = 1$, otherwise $A_{i,j} = 0$, since G is an undirected graph, the adjacency matrix A is a symmetric matrix. A sequence of human action diagrams is represented as a set of joint point feature maps $X = \{x_{c,t,n} | c = 1, 2, ..., C; t = 1, 2, ..., T; n = 1, 2, ..., N\}$, where C is the number of dimensions of the feature, T is the number of frames of the skeleton sequence, and N is the total number of human joints in a frame. So that the input action is adequately described structurally by A and feature-wise by X.

3.1.2 Graph convolution network

In ST-GCN, the adjacency matrix of the graph divides the root node and its neighbor nodes into three parts according to the division strategy: (1) The graph A_1 composed of the root node itself. (2) The graph A_2 composed of centripetal nodes, the neighbor nodes closer to the center of gravity of the skeleton than the root node. (3) The other neighboring nodes are centrifugal nodes and comprise the graph A_3. For the skeleton input defined by the feature map X, the formula of ST-GCN for performing the graph convolution can be represented as:

$$X^{(l+1)} = \sigma(\sum_{k=1}^{3} D_k^{-\frac{1}{2}} A_k D_k^{-\frac{1}{2}} X^{(l)} W_k{}^{(l)}) \tag{1}$$

Where D_k is the degree matrix of A_k, $X^{(l)}$ is the feature map of the layer l. $W_k{}^{(l)}$ represents the weight matrix of layer l in a convolution network, which is a learnable parameter, and σ is the activation function.

3.2 Enhanced decoupling graph convolution network

In order to increase the diversity of input features and make the model learn more diverse and rich features from the skeleton sequence, we adopts a framework of multi-stream fusion. Since both visualized joints and bones can help humans judge actions, the preprocessed input features in this paper are divided into two main categories: joint stream and bone stream. The ED-GCN network is trained separately using different input features, and the output classification results are fused.

The overview of the proposed ED-GCN is shown in Fig. 1. ED-GCN stacks 10 basic blocks together, each including a serially connected ED-GC module, a simplified multi-stage temporal convolutional network (MS-TCN) module and a STA module. The output channels of each block are 64, 64, 64, 64, 128, 128, 128, 256, 256, 256. ED-GCN aggregates features using a global average pooling layer at the end of the network, and uses a fully connected layer for classification. To avoid overfitting, a dropout layer is added before the final fully connected layer and set to 0.4. The dropout layer can cause some neurons to stop working randomly while the network is being trained. All activations in our model use the Mish activation function [33], which is similar to the ReLU function, but is smoother than it.

3.2.1 Enhanced decoupling graph convolution module

The utilization of a predefined adjacency graph limits itself to considering connections between individual nodes and their immediate neighbors. Consequently, this results in closer nodes and nodes with higher degrees carrying more weight in the network learning process. As a consequence, a significant portion of learned information becomes concentrated within these nodes. Meanwhile, connections between distant nodes necessitate multiple layers of graph convolutions to be stacked, leading to inefficiency in learning these relationships. In order to appropriately expand the perceptual field of graph convolution, we add two additional graphs to the three graphs proposed by ST-GCN: (1) graph A_4 consisting of 2-hop edges. (2) graph A_5 consisting of 3-hop edges.

The graph convolution method proposed in ST-GCN makes all channels of the feature map share the same adjacency matrix, and the graph convolution aggregates features of the same topology among different channels. However, each channel of the feature map represents a different type of feature, so its expressiveness will be limited.

To better optimize the topology represented in each adjacency matrix, a trainable attention matrix $B_k \in \mathbb{R}^{4 \times N \times N}$ is used multiplied with a trainable scale factor α, which is then

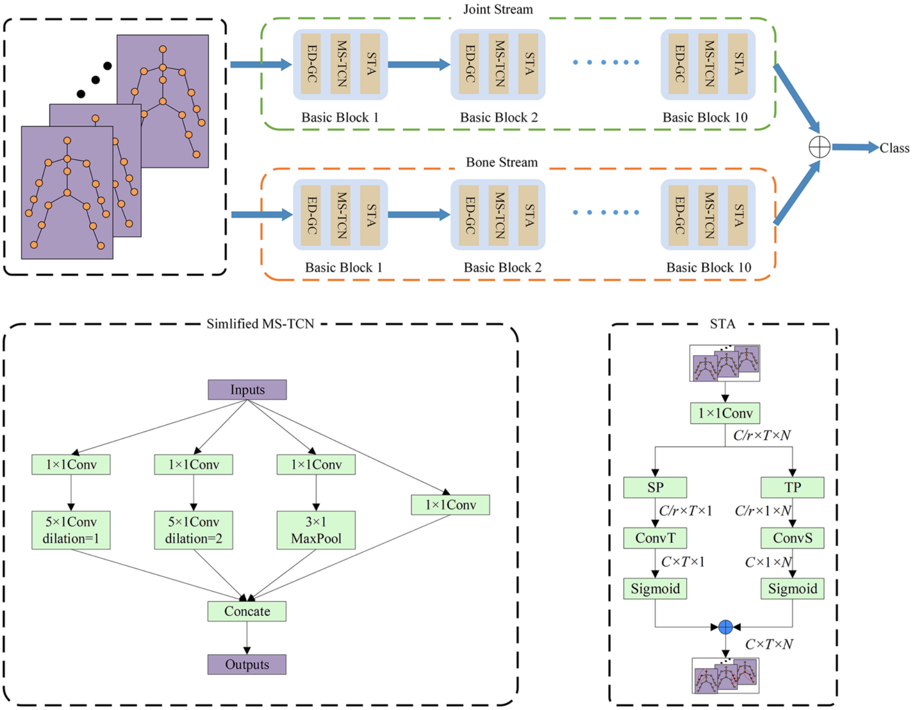

Fig. 1 An overview of the proposed ED-GCN model. For the STA module, r = 2 is utilized to compact the features, where SP denotes spatial pooling, TP denotes temporal pooling, ConvS denotes the convolution along spatial dimension, ConvT denotes the convolution long temporal dimension

combined with the adjacency matrix. This process is expressed using the formula as follows:

$$\tilde{A}_k = D_k{}^{-\frac{1}{2}} A_k D_k{}^{-\frac{1}{2}} + \alpha B_k \tag{2}$$

were $\tilde{A}_k \in \mathbb{R}^{4\times N\times N}$, B_k is initialized to some random value close to 0 as a trainable parameter that can be optimized along with the network, adjusting the strength of the connections between joints defined in the matrix A_k, not only to create new edges but also to remove old ones. The scale factor α is initialized to 0 and is learned entirely from the data, controlling the effect of the attention matrix B_k on the overall graph convolution operation. It makes the graph convolution process more flexible.

The scale factor α plays a crucial role in controlling the impact of the attention matrix B_k on the overall graph convolution operation. At the outset, it is initialized to 0. As training progresses, α is modified based on the model's understanding of the task and the relationships between nodes in the graph. This adjustment enables the model to effectively determine how much attention-based information should be integrated alongside the structural information during the graph convolution operation.

The scale factor α acts as a weighting factor that balances the contributions of the attention matrix B_k and the degree matrix A_k in the graph convolution operation. A smaller value of α puts more emphasis on the structural relationships captured by A_k, essentially relying on the inherent graph topology. Conversely, as α increases, the attention matrix B_k gains more importance, enabling the model to focus on finer relationships indicated by the attention mechanism. The flexibility of α allows the model to dynamically shift its focus between the two sources of information based on the data's characteristics and the learning objectives.

In summary, the scale factor α is not static; it is learned, adjusted, and fine-tuned as the model trains on data. Its value determines the extent to which attention-based connections influence the graph convolution process in conjunction with the structural information. This adaptability empowers the model to effectively capture both local and global dependencies, ultimately enhancing its ability to extract meaningful patterns from the data.

We divide the channels of the whole feature map into four groups, and each group of channels has a separate learnable adjacency matrix. The process of the ED-GC module is expressed using the formula as follows:

$$X^{(l+1)} = \sigma(\sum_{k=1}^{5}(\tilde{A}_k^1 {X^{(l)}}_{0:\frac{C}{4},:,:} \| \tilde{A}_k^2 {X^{(l)}}_{\frac{C}{4}:\frac{C}{2},:,:} \| \tilde{A}_k^3 {X^{(l)}}_{\frac{C}{2}:\frac{3C}{4},:,:} \| \tilde{A}_k^4 {X^{(l)}}_{\frac{3C}{4}:C,:,:}) {W_k}^{(l)}) \tag{3}$$

Where $\tilde{A}_k^i \in \mathbb{R}^{N \times N}$ represents the subset i of $\tilde{A}_k$, $\|$ represents the channel-wise concatenation, σ represents the activation function. ${W_k}^{(l)}$ represents the learnable weight matrix of layer l.

3.2.2 Temporal convolution module

The temporal information is crucial for continuous video frames because it include a lot of associate information, particularly for analyses like action recognition. The performance of the prior temporal convolution module, which merely used basic temporal convolution to extract the temporal information from activities, has been subpar.

In [27], it is suggested to use the MS-TCN module to collect temporal information. This method has achieved outstanding performance, However, its excessive branching reduces the speed of inference.

We reduced on some dilated convolution branches while increasing the size of convolution kernel to increase receptive field. The capacity of MS-TCN to extract information is preserved while the complexity of model are significantly reduced. The same fantastic results are obtained with this procedure. The overview of the simplified MS-TCN module is shown in Fig. 1.

3.2.3 Spatial-temporal union attention module

The overview of the proposed STA module is shown in Fig. 1. In order to reduce the number of parameters and computation of the model, the STA module first compresses the channel dimension of the feature map using 1×1 convolution, then the temporal attention branch compresses the spatial dimension of the feature map using average pooling, then extracts the temporal attention information in the temporal dimension using convolution, and normalizes the temporal attention information after the Sigmoid activation function, and the importance of temporal features is scored to derive the score of temporal attention information within each channel. The spatial attention branch uses a similar method to extract the spatial attention information of the feature map. Finally, the outputs of these two branches are summed to obtain the attention scores of each joint in the whole action sequence. The process of processing data by the STA module can be expressed in the following formula:

$$X' = g(X) \tag{4}$$

$$X'' = \sigma(g_t(pool_s(X'))) + \sigma(g_s(pool_t(X'))) \tag{5}$$

$$X_{out} = X \odot X'' \tag{6}$$

where $pool_s$ denotes the average pooling operation in the spatial dimension, $pool_t$ denotes the average pooling operation in the temporal dimension, g_t denotes the convolution operation along the temporal dimension, g_s denotes the convolution operation along the spatial dimension, σ denotes the Sigmoid activation function, g denotes the 1×1 convolution operation, and X_{out} denotes the final output feature map incorporating spatial-temporal attention information.

4 Experiments

4.1 Datasets

4.1.1 NTU RGB+D 60

The most reputable and popular large skeleton-based action recognition dataset is NTU RGB+D 60 [11]. It has a total of 56,880 skeleton sequences and 40 distinct people. Each video is taken using three cameras at the same height but with various horizontal angles: -45°, 0°, and 45°. The KinectV2 depth sensor is used to determine the 3D joint position coordinates of the human body for each frame. There are 60 different action classes. Each skeleton sequence has a maximum of 2 subjects, each with 25 joints. Two protocols are suggested by the authors of the study that proposes the dataset: (1) In cross-subject (X-Sub), 40 subjects are splited into a training group and a test group depending on their IDs, and including 40320 samples in the training group and 16560 samples in the test group. (2) In cross-view (X-View), 18960 samples from camera 1 are used as the test group, and 37920 samples from cameras 2 and 3 as the training group.

4.1.2 NTU RGB+D120

The biggest skeleton-based large action recognition dataset currently available is NTU RGB+D120 [12], which is an upgraded version of NTU RGB+D 60. There are 120 action classes, 114,480 skeleton sequences, 155 camera viewpoints, 96 distinct backdrops, and data from 106 people in this dataset. The people were from 15 different nations, with ages ranging from 10 to 57 and heights varying from 1.3 to 1.9 meters. As can be seen, there is a lot of variance across the participants, which further improves the dataset's quality. Two protocols are suggested by the authors of the study that proposes the dataset: (1) In cross subjects (X-Sub120), 106 subjects are separated the into two groups depending on their IDs, one for training and one for testing, with the training group including 63026 samples and the test group containing 51454 samples. (2) In cross setting (X-Set120), 54471 samples are used as the training group and 60009 samples as the test group, with the groups being split according to the people' locations and backgrounds.

4.1.3 Northwestern-UCLA

Three Kinect cameras are used to record the Northwestern-UCLA dataset. It has 1494 video clips in 10 different categories. There are ten actors involved in each action. We follow the protocol in [13] for this analysis, using the samples from the first two cameras as training data and the samples from the second camera as testing data.

4.2 Implementation details

To verify the advancement and the efficiency of ED-GCN, the ED-GCN model is compared with the state-of-the-art methods on the NTU RGB+D 60, NTU RGB+D 120, and Northwestern-UCLA datasets, respectively. All experiments were conducted on two RTX3090 GPUs using the Pytorch deep learning framework [34], and our models used an SGD function with momentum of 0.9 and weight decay of 0.0001 as the optimization strategy. The learning rate is set to 0.1 and multiplied by a factor of 0.1 for epochs 20 and 30. For NTU RGB+D 60 and NTU RGB+D 120, with a total of 80 training epochs, we sample the data from each sample so that each sample contains 100 frames and the batch size is set to 32. For the X-View protocol, we use the data preprocessing method [31]. For Northwestern-UCLA, with a total of 200 training epochs and batch size set to 16.

Table 1 Classification accuracy comparison against state-of-the-art methods on the NTU RGB+D 60 dataset

Methods	Accuracy (%)	
	X-Sub	X-View
ST-GCN [10]	81.5	88.3
AS-GCN [26]	86.8	94.2
SGN [31]	89.0	94.5
PA-ResGCN-B19 [35]	**90.9**	**96.0**
ED-GCN (Joint only, ours)	90.8	95.7
ED-GCN (Bone only, ours)	90.2	95.1
3s-AdaSGN [36]	90.5	95.3
4s-Shift-GCN [32]	90.7	96.5
DC-GCN+ADG [37]	90.8	96.6
MST-GCN [38]	91.5	96.6
InfoGCN (6 ensemble) [29]	**93.0**	**97.1**
2s-AGCN [23]	88.5	95.1
SATD-GCN [28]	89.3	95.5
DGNN [25]	89.9	96.1
MS-AAGCN [24]	90.0	96.2
MS-G3D [27]	91.5	96.2
ED-GCN (ours)	**91.6**	**96.5**

X-Sub and X-Set represent cross-subject and cross-view, respectively
The bold entries represent the best results in each group

Table 2 Classification accuracy comparison against state-of-the-art methods on the NTU RGB+D 120 dataset

Methods	Accuracy (%)	
	X-Sub120	X-View120
AS-GCN [26]	77.9	78.5
SGN [31]	79.2	81.5
PA-ResGCN-B19 [35]	**87.3**	**88.3**
ED-GCN (Joint only, ours)	86.3	87.8
ED-GCN (Bone only, ours)	86.1	87.3
3s-AdaSGN [36]	85.9	86.8
4s-Shift-GCN [32]	85.9	87.6
DC-GCN+ADG [37]	86.5	88.1
MST-GCN [38]	87.5	88.8
InfoGCN (6 ensemble) [29]	**89.8**	**91.2**
2s-AGCN [23]	82.9	84.9
MS-G3D [27]	86.9	88.4
InfoGCN (Joint+Bone) [29]	**88.5**	89.7
ED-GCN (ours)	88.2	**90.0**

X-Sub120 and X-Set120 represent cross-subject and cross-setup splits, respectively
The bold entries represent the best results in each group

4.3 Comparisons with the State-of-the-Arts

The comparison results are listed in following tables, respectively. Tables 1, 2 and 4 are divided into three parts based on the number of streams utilized by the models. The top part features models employing a single stream, the bottom part showcases models with two streams, and the middle part presents models integrating more than two streams.

From Table 1, it can be observed that ED-GCN achieves 91.6% and 96.5% classification accuracy on the two protocols of the NTU RGB+D 60 dataset, respectively. Compared with the most popular backbone model for skeleton-based action recognition, i.e. ST-GCN, ED-GCN exceeds it by 10.1% and 8.2% on two protocols, respectively. Moreover, ED-GCN performs better than all of 1-stream and 2-stream models. It should be noticed that DC-GCN+ADG and MST-GCN achieve slightly higher accuracies than ours on the X-View protocol, and the performance of InfoGCN is also ahead of our on two protocols. These methods have a common characteristic, which is the strategy of fusing more than two streams. This strategy fuses additional motion information, which may enhance the representation of cross-view skeleton features. The superior performance of these models is achieved at the expense of model complexity and computational costs.

From Tables 2 and 3, it can be observed that ED-GCN achieves classification accuracies of 88.2% and 90.0% on the two protocols of the NTU RGB+D 120 dataset, and an accuracy of 95.1% on the Northwestern-UCLA dataset. These performances outperform all the models in the two tables, except for InfoGCN (6 ensemble). Surprisingly, our model is superior to InfoGCN (Joint+Bone) on the X-Set120 protocol of NTU RGB+D 120 dataset. It is indicated that our model is not inferior to InfoGCN, if they all utilize two streams (Joint+Bone).

To verify the efficiency of ED-GCN, we also compare the model complexity in terms of FLOPs and number of parameters (# Param.) on the X-View protocol of the NTU RGB+D 60 dataset. The results are listed in Table 4. Since the FLOPs and the parameter numbers of

Table 3 Classification accuracy comparison against state-of-the-art methods on the Northwestern-UCLA dataset

Methods	Accuracy (%)
Lie Group [39]	74.2
HBRNN-L [1]	78.5
Ensemble TS-LSTM [40]	89.2
AGC-LSTM [41]	93.3
4s-Shift-GCN [32]	94.6
InforGCN (6 ensemble) [29]	**97.0**
ED-GCN (Joint only, ours)	94.6
ED-GCN (Bone only, ours)	93.8
ED-GCN (ours)	95.1

The bold entries represent the best results

some models are not found, and the size of input skeleton sequence of SGN and 3s-AdaSGN is different from other models, the FLOPs and the parameter numbers of these models are not presented in this table. It is evident that our proposed ED-GCN exhibits the lowest FLOPs while sharing a parameter count similar to that of 4s-Shift-GCN, yet remaining smaller than the other models. Moreover, ED-GCN is about 1.78x faster and 3.30x smaller than InfoGCN (6 ensemble) with the SOTA accurauacy. The results show that our proposed ED-GCN model is a powerful and efficient model in the field of skeleton-based action recognition.

Table 4 Model complexity comparison against state-of-the-art methods on X-Sub of NTU RGB+D 60 dataset

Methods	FLOPs ($\times 10^9$)	# Param. ($\times 10^6$)
ST-GCN [10]	**16.32**	**3.10**
AS-GCN [26]	26.76	9.50
SGN [31]	-	-
PA-ResGCN-B19 [35]	18.52	3.64
3s-AdaSGN [36]	-	-
4s-Shift-GCN [32]	**10.00**	**2.76**
DC-GCN+ADG [37]	25.72	4.96
MST-GCN [38]	-	12.00
InfoGCN (6 ensemble) [29]	15.50	9.22
2s-AGCN [23]	37.32	6.94
SATD-GCN [28]	-	-
DGNN [25]	-	26.24
MS-AAGCN [24]	-	-
MS-G3D [27]	48.88	6.40
ED-GCN (ours)	**8.71**	**2.79**

The bold entries represent the best results in each group

Table 5 The comparison of model accuracy under different settings

Settings	Accuracy (%)
ST-GCN (baseline)	81.5
ED-GC (w/o $A_4\&A_5$)	90.0
ED-GC	90.8

w/o means delete this part

4.4 Ablation experiments

In this section, we analyze the contributions of different components in the proposed ED-GCN, including the ED-GC module and the STA module. We choose the ST-GCN as the baseline model. All experiments are studied on the NTU RGB+D 60 dataset using a single joint stream on the X-Sub protocol.

4.4.1 Effectiveness of ED-GC module

The ED-GC module contains two extra graphs, namely A4 and A5, to expand the perceptual field. To evaluate the impact of these additional graphs on the model, we conducted separate tests with and without their inclusion, and the corresponding results are summarized in Table 5. The findings demonstrate that the absence of these two additional graphs led to a drop in accuracy to 90.0%, underscoring the significant improvement in action recognition accuracy due to their incorporation. Furthermore, our model's accuracy surpasses that of the baseline significantly, further affirming the effectiveness of the ED-GC module.

In traditional graph convolution, a shared spatial aggregation kernel is the norm, typically embodied by an adjacency matrix. However, this convention imposes a constraint on the potential depth of spatial aggregation. While augmenting the number of adjacency matrices can partly alleviate this concern, it unavoidably leads to a proportional surge in computational demands, thus compromising efficiency. Enter ED-GC module, where the channel-wise features undergo a clever division into four distinct groups, each bestowed with its own trainable adjacency graph. This ingenious approach imbues spatial aggregation with heightened expressiveness. Notably, ED-GC module introduces a dual cadre of connections, which signify long-range nodal associations, effectively broadening the network's receptive field. This strategic augmentation of connectivity remarkably enhances spatial aggregation's expressiveness, all the while sidestepping any substantial escalation in computational overhead.

4.4.2 Effectiveness of STA module

To explore the effectiveness of the STA module and its impact on the model, we removed this module from the model and tested it. In Table 6, it can be observed that the addition of the STA module has improved the accuracy of the model, and it has improved the accuracy of the model by 1.1% and brought the model to the highest accuracy of 90.8%. The STA module

Table 6 Impact of STA module on model accuracy

Settings	Accuracy (%)
ED-GCN (w/o STA)	89.7
ED-GCN	90.8

Table 7 Comparison of multi-stream and single-stream methods

Settings	Accuracy (%)
Joint	90.8
Bone	90.2
Joint+ Bone	91.6

performs spatial and temporal attention in parallel, and then fuses the attention information to obtain a global spatio-temporal attention feature, which is finally multiplied with the feature maps. This not only effectively considers the correlation between spatial and temporal attention but also greatly reduces the computational cost and improves the inference speed of the model. Given that ED-GCN groups features along the channels and requires a large number of learned adjacency matrices, it is necessary for the attention module to extract spatial feature importance information to enable the model to emphasize the connection between distant nodes with strong correlation. At the same time, learning temporal attention information allows the model to consider the spatiotemporal attention correlation, utilize limited parameters to focus on learning important frames in the features, and achieve better results. Therefore, the STA module is very important for the ED-GCN network.

4.4.3 Effectiveness of multi-stream network

In order to test the performance of the proposed two streams, experiments were conducted and the results are shown in Table 7, where Joint denotes the joint stream and Bone denotes the bone stream. It can be observed that the performance of the model is relatively low when only a single joint stream is used, and the best performance of the ED-GCN model is obtained when both streams are superimposed and used together, each of which allows the model to learn different aspects of information effectively, and the multi-stream fusion approach is clearly superior to the single-stream approach and can effectively improve the performance of the model.

On the X-Sub protocol of NTU RGB+D 60 dataset, we compared the accuracy of ED-GCN and the baseline model ST-GCN on each category, and the results are shown in Fig. 2. As can be seen, our proposed ED-GCN performs much better than the baseline and even increases accuracy by more than 20% in categories 10,11,12,29,31,32 and 33.

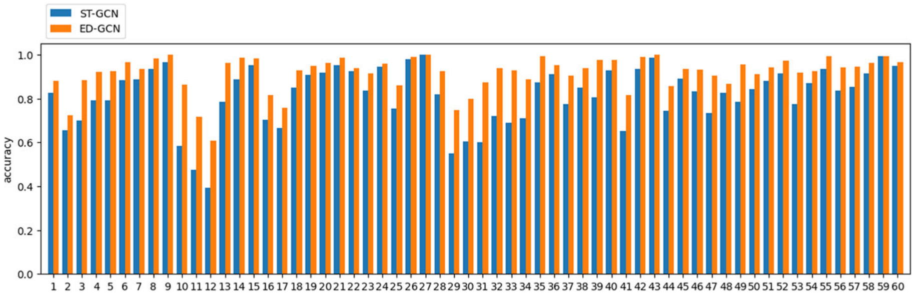

Fig. 2 Comparison of the classification accuracy of ED-GCN and ST-GCN for each category on the NTU RGB+D 60 dataset. We use numbers to denote the name of each category

5 Conclusion

In this paper, a novel enhanced decoupling graph convolution network is proposed for skeleton-based action recognition. the ED-GC module effectively expands the perceptual field of the graph convolution, increasing its expressive power and flexibility. In addition, the STA module is embedded into each basic block, while finding and fusing important spatio-temporal features in the whole skeleton sequence from both temporal and spatial perspectives, embedding attention information into the model, thus helping the model to focus on important elements and extract features more efficiently. Extensive experiments on three datasets demonstrate the effectiveness and advancedness of the ED-GCN model in skeleton-based action recognition.

Acknowledgements This research was supported by the National Natural Science Foundation ofChina (61806145, 62376197, 61906135, 62020106004, 92048301) and the Natural Science Foundation of Tianjin City (21JCQNJC01510).

Data Availability The NTU RGB+D 60, NTU RGB+D120 and Northwestern-UCLA dataset are openly available at https://rose1.ntu.edu.sg/dataset/actionRecognition/ and http://users.eecs.northwestern.edu/jwa368/my_data.html.

Declarations

Conflict of interest The authors declare that there is no conflict of interest that would affect the work of this paper.

References

1. Du Y, Wang W, Wang L (2015) Hierarchical recurrent neural network for skeleton based action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1110–1118
2. Veeriah V, Zhuang N, Qi G-J (2015) Differential recurrent neural networks for action recognition. In: Proceedings of the IEEE international conference on computer vision, pp 4041–4049
3. Wang H, Wang L (2017) Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 499–508
4. Liu J, Shahroudy A, Xu D, Wang G (2016) Spatio-temporal lstm with trust gates for 3d human action recognition. In: European conference on computer vision, pp 816–833
5. Liu J, Wang G, Hu P, Duan L-Y, Kot AC (2017) Global context-aware attention lstm networks for 3d action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1647–1656
6. Li C, Hou Y, Wang P, Li W (2017) Joint distance maps based action recognition with convolutional neural networks. IEEE Signal Process Lett 24(5):624–628
7. Wang P, Li Z, Hou Y, Li W (2016) Action recognition based on joint trajectory maps using convolutional neural networks. In: Proceedings of the 24th ACM international conference on multimedia, pp 102–106
8. Soo Kim T, Reiter A (2017) Interpretable 3d human action analysis with temporal convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 20–28
9. Li C, Zhong Q, Xie D, Pu S (2018) Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. arXiv:1804.06055
10. Yan S, Xiong Y, Lin D (2018) Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-second AAAI conference on artificial intelligence
11. Shahroudy A, Liu J, Ng T-T, Wang G (2016) Ntu rgb+ d: a large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1010–1019
12. Liu J, Shahroudy A, Perez M, Wang G, Duan L-Y, Kot AC (2019) Ntu rgb+ d 120: a large-scale benchmark for 3d human activity understanding. IEEE Trans Pattern Anal Mach Intell 42(10):2684–2701

13. Wang J, Nie X, Xia Y, Wu Y, Zhu S-C (2014) Cross-view action modeling, learning and recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2649–2656
14. Li R, Wang S, Zhu F, Huang J (2018) Adaptive graph convolutional neural networks. In: Proceedings of the AAAI conference on artificial intelligence, vol 32
15. Defferrard M, Bresson X, Vandergheynst P (2016) Convolutional neural networks on graphs with fast localized spectral filtering. Adv Neural Inf Process Syst 29
16. Atwood J, Towsley D (2016) Diffusion-convolutional neural networks. Adv Neural Inf Process Syst 29
17. Xu K, Hu W, Leskovec J, Jegelka S (2018) How powerful are graph neural networks? arXiv:1810.00826
18. Hamilton W, Ying Z, Leskovec J (2017) Inductive representation learning on large graphs. Adv Neural Inf Process Syst 30
19. Abu-El-Haija S, Perozzi B, Kapoor A, Alipourfard N, Lerman K, Harutyunyan H, Ver Steeg G, Galstyan A (2019) Mixhop: higher-order graph convolutional architectures via sparsified neighborhood mixing. In: International conference on machine learning, pp 21–29
20. Kipf TN, Welling M (2016) Semi-supervised classification with graph convolutional networks. arXiv:1609.02907
21. Bruna J, Zaremba W, Szlam A, LeCun Y (2013) Spectral networks and locally connected networks on graphs. arXiv:1312.6203
22. Hammond DK, Vandergheynst P, Gribonval R (2011) Wavelets on graphs via spectral graph theory. Appl Comput Harmonic Anal 30(2):129–150
23. Shi L, Zhang Y, Cheng J, Lu H (2019) Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12026–12035
24. Shi L, Zhang Y, Cheng J, Lu H (2020) Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. IEEE Trans Image Process 29:9532–9545
25. Shi L, Zhang Y, Cheng J, Lu H (2019) Skeleton-based action recognition with directed graph neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7912–7921
26. Li M, Chen S, Chen X, Zhang Y, Wang Y, Tian Q (2019) Actional-structural graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3595–3603
27. Liu Z, Zhang H, Chen Z, Wang Z, Ouyang W (2020) Disentangling and unifying graph convolutions for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 143–152
28. Zhang J, Ye G, Tu Z, Qin Y, Qin Q, Zhang J, Liu J (2022) A spatial attentive and temporal dilated (satd) gcn for skeleton-based action recognition. CAAI Trans Intell Technol 7(1):46–55
29. Chi H-g, Ha MH, Chi S, Lee SW, Huang Q, Ramani K (2022) Infogcn: representation learning for human skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 20186–20196
30. Duan H, Zhao Y, Chen K, Lin D, Dai B (2022) Revisiting skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2969–2978
31. Zhang P, Lan C, Zeng W, Xing J, Xue J, Zheng N (2020) Semantics-guided neural networks for efficient skeleton-based human action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1112–1121
32. Cheng K, Zhang Y, He X, Chen W, Cheng J, Lu H (2020) Skeleton-based action recognition with shift graph convolutional network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 183–192
33. Misra D (2019) Mish: a self regularized non-monotonic activation function. arXiv:1908.08681
34. 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
35. Song Y-F, Zhang Z, Shan C, Wang L (2020) Stronger, faster and more explainable: a graph convolutional baseline for skeleton-based action recognition. In: Proceedings of the 28th ACM international conference on multimedia, pp 1625–1633
36. Shi L, Zhang Y, Cheng J, Lu H (2021) Adasgn: adapting joint number and model size for efficient skeleton-based action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 13413–13422
37. Cheng K, Zhang Y, Cao C, Shi L, Cheng J, Lu H (2020) Decoupling gcn with dropgraph module for skeleton-based action recognition. In: European conference on computer vision, pp 536–553
38. Chen Z, Li S, Yang B, Li Q, Liu H (2021) Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition. Proceedings of the AAAI Conference on Artificial Intelligence 35:1113–1122

39. Vemulapalli R, Arrate F, Chellappa R (2014) Human action recognition by representing 3d skeletons as points in a lie group. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 588–595
40. Lee I, Kim D, Kang S, Lee S (2017) Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks. In: Proceedings of the IEEE international conference on computer vision, pp 1012–1020
41. Si C, Chen W, Wang W, Wang L, Tan T (2019) An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1227–1236

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.