Article

Research on Multi-Scale Spatio-Temporal Graph Convolutional Human Behavior Recognition Method Incorporating Multi-Granularity Features

Yulin Wang , Tao Song * , Yichen Yang , Zheng Hong

Posted Date: 15 October 2024

doi: 10.20944/preprints202410.1179.v1

Keywords: graph convolutional network; behavior recognition; multiscale; bias weighting

Article

Research on Multi-Scale Spatio-Temporal Graph Convolutional Human Behavior Recognition Method Incorporating Multi-Granularity Features

Yulin Wang [1], Tao Song [2,*], Yichen Yang [2] and Zheng Hong [1]

[1] Chongqing Vocational College of Public Transportation, Chongqing 402260, China
[2] Chongqing University of Technology, Chongqing 400054, China
* Correspondence: tsong@cqut.edu.cn

Abstract: Aiming at the problem that the existing human skeleton behavior recognition methods are insensitive to human local movements and have inaccurate recognition in distinguishing similar behaviors, a multi-scale spatio-temporal graph convolution method incorporating multi-granularity features is proposed for human behavior recognition. Firstly, a skeleton fine-grained division strategy is proposed, which initializes the skeleton data into data streams of different granularities. Using a normalized Gaussian function, an adaptive cross scale feature fusion layer is designed for feature fusion between different granularities, and fine-grained features guide the model to focus on discriminative feature expressions between similar behaviors. Secondly, a sparse multi-scale adjacency matrix is introduced to solve the bias weighting problem that amplifies the multi-scale spatial domain modeling process under multi granularity conditions. Finally, an end-to-end graph convolutional neural network is constructed to improve the feature expression ability of spatio-temporal receptive field information and enhance the robustness of recognition between similar behaviors. The feasibility of the proposed algorithm was verified on the public behavior recognition dataset MSR Action 3D, with a recognition rate of 95.67%, which is superior to most existing behavior recognition methods.

Keywords: graph convolutional network; behavior recognition; multiscale; bias weighting

1. Introduction

As an extremely important component of the computer vision field, research on behavior recognition has always been of great concern and widely applied, with broad application prospects in intelligent monitoring, motion analysis, human-computer interaction, and other fields [1–3]. At present, human skeleton behavior recognition based on deep learning is mainly divided into three categories: The first type is to use Convolutional Neural Networks [4–6] to model skeleton data as pseudo images, extracting highly abstract skeletal structural features. The second type is to use Recurrent Neural Networks [7–10] to model skeleton data as sequences of coordinate vectors, capturing the dynamic correlations between consecutive frames of skeletal data to predict behavior categories. The last type is Graph Convolutional Network (GCN), which represents the human skeleton sequence as a spatio-temporal topological graph, by using graph convolution, the global features of the skeleton spatial structure are effectively extracted, which can better model the spatio-temporal characteristics of human skeleton information. Therefore, graph convolution based human skeleton behavior recognition methods have become a research hotspot in recent years.

Reference [11] proposed a Spatio-Temporal Graph Convolutional Network (ST-GCN), which was used for the first time to model human skeleton data in a spatio-temporal graph and achieved good recognition results. Shi et al. [12] proposed adaptive graph convolution, which calculates the similarity between joint points based on input skeleton data of different action classes to adaptively measure the degree of correlation between joint points. Li et al. [13] proposed a motion structure that emphasizes the dependency relationship between non-adjacent joint points in space through action linking modules and structural linking modules. References [14–16] proposed a multi-scale spatial graph convolutional network to capture feature information between nodes in a wider space, using high-order polynomials of the adjacency matrix to aggregate features between remote nodes. However, these methods have bias weighting issues in the process of spatial domain modeling, which means that in the process of modeling spatial position relationships using high-order adjacency matrices, nodes far from the target joint point make little contribution to recognition, and the final recognition result will be dominated by joints from local body parts. Meanwhile, due to the presence of information from different modalities and spatio-temporal scales within the skeleton, all of which are crucial for behavior recognition, many works have attempted to explore and utilize this information. Shi et al. [12] added the inter-frame difference between the bone flow and keypoint flow in 2s-AGCN as the information for keypoint motion flow and bone motion flow. The Shift-GCN network proposed by Cheng et al. [17] performs more processing on the original data, extracting frame differences as dynamic information between keypoints and bones based on keypoint coordinates and bone vectors, these four different forms of data are used as inputs to jointly predict category features.

In Li et al.'s [18] study, higher-order transformations were applied to the original skeleton data, subsequently, a multi-stream network was employed for decision-level fusion of advanced information, such as joint and bone details, thereby further improving the performance of the model. However, these methods do not take into account the spatial granularity featrues of human behavior processes, from the perspective of human kinematics, the recognition of certain behavioral actions depends on the characteristics between distant nodes, while the recognition of similar behaviors relies more on subtle motion differences between local nodes.

Therefore, aiming at the above problems, a multi-scale spatio-temporal graph convolution method incorporating multi-granularity features is proposed for human behavior recognition. Initialize the input skeleton data into data streams of different gTherefore, aiming at the above problems, a multi-scale spatio-temporal graph convolution method incorporating multi-granularity features is proposed ranularities to guide the network to learn the differences between similar behaviors, and construct a cross scale fusion module for feature fusion between different granularities; by constructing a multi-scale adjacency matrix and subtracting adjacent adjacency matrices at different spatial scales, a sparse adjacency matrix is constructed to solve the bias weighting problem in the process of multi-scale spatial modeling; an end-to-end multi-scale graph convolution network incorporating multi-granularity features is constructed, and the feasibility of the algorithm is verified on the public behavior recognition dataset MSR Action 3D..

2. Skeleton Behavior Recognition Based on Graph Convolution

2.1. Spatio-Temporal Graph Convolutional Network

GCN is widely used in the modeling of human skeleton data, in this method, the human skeleton is generally represented as a spatio-temporal graph $G = (V, E)$ with N joints and T frames, where V represents the joints of the skeleton and E represents the edges connecting the human joints. The skeleton coordinates of human actions can be expressed as $X \in R^{C\times T\times N}$, where C is the number of channels, T is the number of frames in the video, and N is the number of nodes in the human skeleton. The GCN-based model mainly consists of two parts: spatial graph convolution and temporal convolution.

In the spatial dimension, the feature extraction of any joint point v_{ti} in the skeleton graph by graph convolution operation is expressed as:

$$f_{out}(v_{ti}) = \sum_{v_{tj}\in B(v_{ti})} \frac{1}{Z_{ti}(v_{tj})} f_{in}(\mathbf{p}(v_{ti}, v_{tj})) \cdot \mathbf{w}(v_{ti}, v_{tj}) \quad (1)$$

Where f1 and f2 represent the input and output features respectively; $\mathrm{B(vti)} = \{\mathrm{v_{ti}} | \mathrm{r}(\mathrm{v_{tj}}, \mathrm{v_{ti}}) \in \mathrm{R}\}$ represents the set of neighboring nodes of v_{ti}, and R controls the range of neighboring nodes selected, $Z_{ti}(v_{tj}) = |\{v_{tk} | l_{ti}(v_{tk}) = l_{ti}(v_{tj})\}|$ is the normalization term; w is the weighting function of neighboring joint points.

The operation of graph convolution in the time domain can be extended from graph convolution in the spatial domain, by using parameter Γ as the time range to control the neighbor set, the neighbor set in both spatial and temporal dimensions is given by:

$$B(v_{ti}) = \{v_{qj} | d(v_{tj}, v_{ti}) \leq K, |q - t| \leq \lfloor \Gamma/2 \rfloor\} \tag{2}$$

The corresponding label mapping set for its neighboring nodes is:

$$l_{\mathrm{ST}}(v_{qj}) = l_{ti}(v_{tj}) + (a - t + \Gamma/2)K \tag{3}$$

where $l_{ti}(v_{tj})$ represents the label mapping of v_{ti} the case of a single frame.

Therefore, on the skeleton input defined by feature X and graph structure A, the output of the network after a layer of graph convolution can be represented as:

$$f_{out} = \sigma\left(D^{-\frac{1}{2}}\tilde{A}D^{-\frac{1}{2}}f_{in}W\right) \tag{4}$$

In the formula, $\tilde{A} = A + I$ represents the skeleton graph structure of the human body, and the connection relationship between joints in the skeleton graph is represented by an $N \times N$ adjacency matrix A and an identity matrix I, D is the degree matrix of each joint point, $D^{-1/2}(A + I)D^{-1/2}$ represents the normalized skeleton structure, W represents learnable weight matrix of the network, and σ is the activated linear layer.

2.2. Analysis of Bias Weighting Problem Methods

The existing methods use high-order polynomials of adjacency matrices to aggregate multi-scale spatial structural information at different moments. Based on formula (4), the update rules for high-order matrices are as follows:

$$f_{out} = \sigma\left(\textstyle\sum_{k=0}^{K} D^{-\frac{1}{2}}\tilde{A}^{k}D^{-\frac{1}{2}}f_{in}W\right) \tag{5}$$

Where K is the highest power of the adjacency matrix, and $\tilde{A}^k$ represents the k-th power matrix of $\tilde{A}$.

The K-order adjacency matrix in a graph convolutional network represents K paths between two nodes, Due to the existence of cyclic traversal between nodes, the number of paths with a distance of K to nodes closer to the current node is greater than the number of nodes that are actually K steps away. This leads to a situation where the network assigns greater weights to nodes that are closer in distance during the modeling process, therefore, when conducting multi-scale modeling in the spatial domain, the aggregated features will be dominated by the motion information of local body parts, making it difficult for the network to effectively capture the dependency relationships between nodes that are farther away.

To address the bias weighting issue mentioned above, reference [19] proposes a multi-scale adjacency matrix, the construction method of the adjacency matrix is redefined as follows:

$$[\widetilde{\mathbf{A}}_{(k)}]_{i,j} = \begin{cases} 1 & \text{if} \quad d(v_i, v_j) = k, \\ 1 & \text{if} \quad i = j, \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Where $d(v_i, v_j)$ provides the shortest distance between two nodes $\mathrm{v_i}$ and $\mathrm{v_j}$, different scales of adjacency matrices can be obtained by setting different k values, meanwhile, the K-order adjacency matrix formula can also be calculated using the following formula:

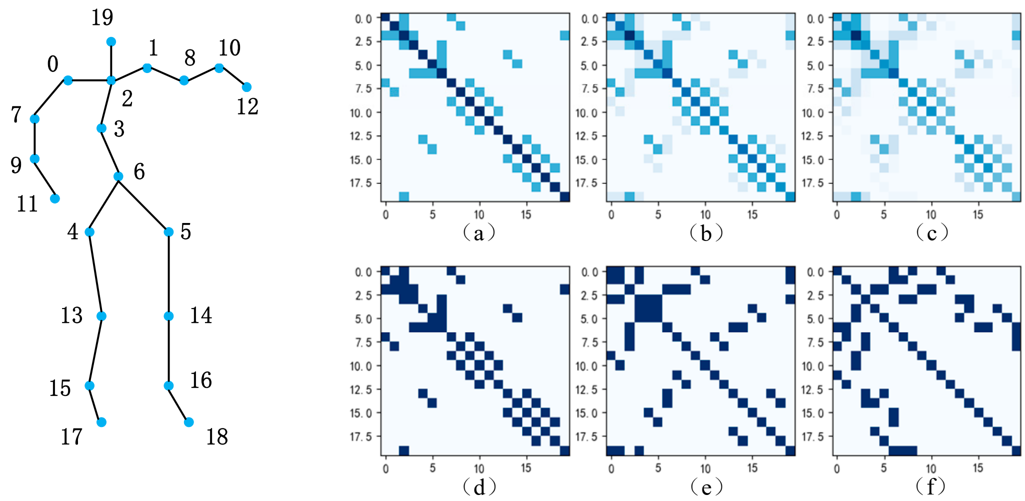
$$\widetilde{\mathbf{A}}_{(k)} = \mathbf{I} + \beth\left(\widetilde{\mathbf{A}}^{k} \geq 1\right) - \beth\left(\widetilde{\mathbf{A}}^{k-1} \geq 1\right) \tag{7}$$

Where $\beth\left(\widetilde{\mathbf{A}}^{k} \geq 1\right)$ represents assigning values greater than or equal to 1 in the matrix to 1, replacing $\widetilde{\mathbf{A}}^{k}$ in equation (5) with $\widetilde{\mathbf{A}}_{(k)}$, we obtain:

$$f_{out} = \sigma\left(\textstyle\sum_{k=0}^{K} D_{(k)}^{-\frac{1}{2}}\widetilde{\mathbf{A}}_{(k)}D_{(k)}^{-\frac{1}{2}}f_{in}W_{(k)}\right) \tag{8}$$

Where $D_{(k)}^{-\frac{1}{2}}\widetilde{\boldsymbol{A}}_{(k)}D_{(k)}^{-\frac{1}{2}}$ represents the standardized K-order adjacency matrix. In this paper, we propose a method of subtracting the K-order adjacency matrix from the K-1 order matrix to eliminate the bias weighting problem that exists in the original modeling approach, which enables the model to better capture the relationship between action categories that are more dependent on features between distant nodes.

As shown in Figure 1, (a), (b), and (c) represent the topological diagrams of the first-order, second-order, and third-order adjacency matrices used to connect human skeletal nodes in a multi-scale spatial model. As the order of the adjacency matrix increases, nodes closer to the current node are assigned greater weights (the darker the color, the greater the weight assigned to the node). Especially when introducing new joint points to refine human skeletal features, the distance between the original two joint points may increase due to the newly added nodes, thus the weight assigned to each other by the two nodes will be further reduced. (d), (e) and (f) represent the topological graphs after constructing a multi-scale adjacency matrix, at this time, the adjacency matrix is reasonably sparsified, allowing the model to assign equal weights to nodes that are farther away, which can better capture the relationships between nodes that are farther away.



（a） （b） （c）

（d） （e） （f）

Figure 1. Adjacency matrix topology diagram.

3. Improved Graph Convolutional Human Behavior Recognition Algorithm

3.1. The Multi-Scale Spatio-Temporal Graph Convolution Network Incorporating Multi-Granularity Features

In order to fully consider the granularity features of human behavior and leverage its advantages in different behavior recognition processes, this paper proposes a multi-scale spatio-temporal graph convolutional network incorporating multi-granularity features for human behavior recognition, the network model framework is shown in Figure 2. Firstly, initialize the joint information of the human body into data streams of different granularity sizes, considering the highly similar behavior categories in the dataset used in this article, it is necessary to refine the joint data to capture more subtle semantic information between behaviors. Secondly, the refined data is fed into the Multi-scale Spatio-temporal Graph Convolutional Block (MS-TGCN) to extract its spatio-temporal features . Then, the obtained output features are fed into the Cross-scale Feature Fusion Layer (CSFL) to blend coarse and fine grained features to capture the differences in features between similar behavior categories. Finally, the fused features are fed into the MS-TGCN layer to further extract its spatio-temporal features, and the prediction results of three granularities are weighted and fused to obtain the classification results.

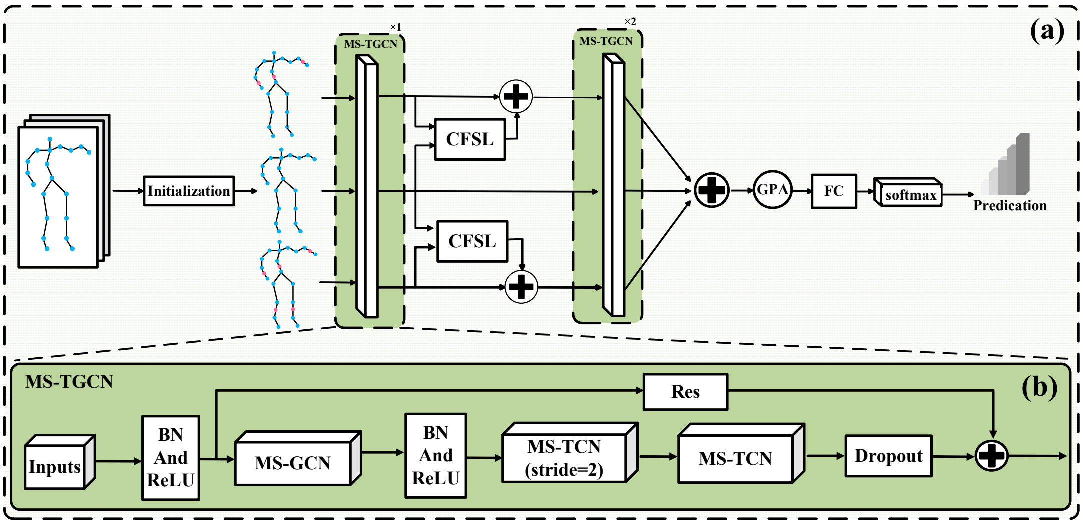


Figure 2. The framework of multi-scale spatio-temporal graph convolutional network model incorporating multi-granularity features.

The model framework of the multi-scale spatio-temporal convolution module is shown in Figure 2b. Firstly, the normalized multi-granularity data stream is fed into MS-TGCN, and the topological relationships between nodes are reconstructed in the spatial domain using a multi-scale adjacency matrix method, by setting different K values, spatial feature fusion is performed on nodes at different distances. Secondly, input the data into two multi-scale time convolutional layers with different step sizes to capture broader temporal contextual features. Finally, the residual module is used to connect the input and output, and the MS-TCN and MS-GCN modules mentioned in reference [19] are used as the multi-scale spatio-temporal graph convolution module in this paper.

3.2. Skeleton Fine-Grained Partitioning Strategy

Due to the high degree of overlap between similar behaviors in the spatio-temporal domain, traditional graph convolutional models are difficult to capture the semantic information that truly distinguishes categories and learn accurate representations. In order to accurately depict fine-grained human behavior, this paper introduces a multi granularity feature learning method, initializing the human skeleton map into different fine-grained levels, as shown in Figure 3. Expand the connections in coarse-grained graphs to tighter connections in fine-grained graphs, enabling fine-grained graphs to represent refined semantic information.

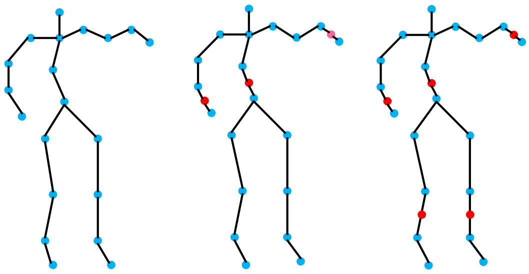


Figure 3. Three granularity representation methods for MSR Action 3D.

The average of multiple adjacent nodes in the coarse-grained graph is calculated through two-dimensional average pooling to represent a supplemented node in the fine-grained scale graph, and then the overall representation of the fine-grained graph is obtained through concatenation. The formula for multi-granularity initialization is expressed as:

$$V_{ck} = \mathrm{pooling}(V_{f1} + V_{f2} + \cdot\cdot + V_{fh}),\mathrm{k} \leqslant \mathrm{h} \tag{9}$$

$$Graph_{\mathrm{new}} = \mathrm{concate}(V_{\mathrm{c1}}, V_{\mathrm{c2}}, \cdots, V_{\mathrm{ck}}) \tag{10}$$

Where V_{ck} represents the joint information of k supplemented nodes in the fine-grained graph, V_{fh} represents the joint information of h nodes in the fine-grained graph, and $Graph_{\mathrm{new}}$ represents the physical skeleton of the fine-grained graph.

3.3. Cross-Scale Feature Fusion Layer

To achieve feature fusion between coarse and fine granularity, fine-grained features are used to guide the original granularity features to learn discriminative feature expressions between similar behaviors, inspired by reference [12], this paper proposes an adaptive cross-scale feature fusion module, as shown in Figure 4.

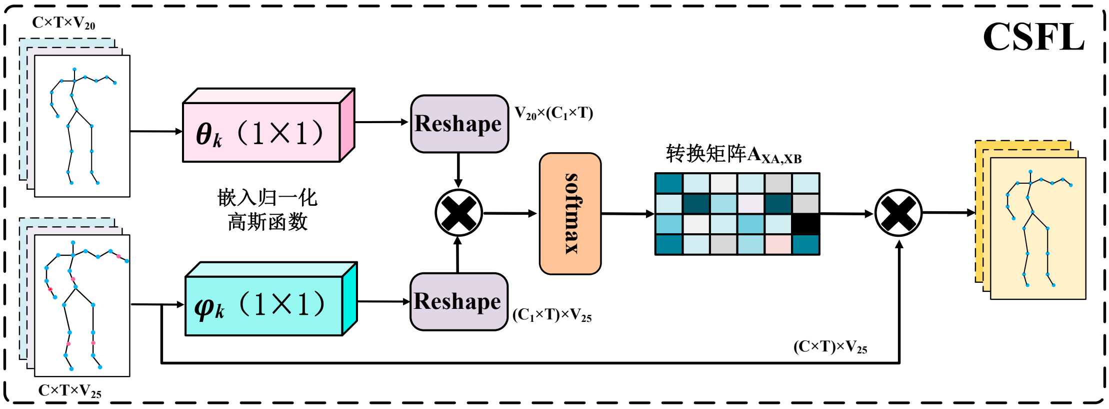


Figure 4. Cross-scale feature fusion layer.

Namely, embedding a normalized Gaussian function in the network to calculate the feature mapping relationship between two sizes and generate a cross-scale feature fusion matrix. The specific operation is as follows:

$$f(v_i, v_j) = \exp[\psi^T(v_i)\theta(v_j)] \,/\, \sum_{j=1}^{N} \exp\,[\psi^T(v_i)\theta(v_j)] \tag{11}$$

Where $\psi^T(v_i) = W_\psi v_i$ and $\theta(v_j) = W_\theta v_j$ represent embedded operations, while W_φ and W_θ are corresponding weight parameters.

Taking 20 coarse-grained skeleton information and 25 fine-grained skeleton information as examples, the feature dimensions of the two input granularity features f_1 and f_2 are $C \times T \times V_{20}$ and $C \times T \times V_{25}$ respectively, where C represents the number of channels of the embedded Gaussian function. The two streams of data undergo a 1×1 convolution operation separately, after performing a dimension transformation on both, matrix multiplication is applied, and finally, an adaptive transformation matrix is obtained through a Softmax classifier as follows:

$$A_{f_1,f_2} = softmax(f_1^T W_\psi^T W_\theta f_2) \in [0,1] \tag{12}$$

This adaptive transformation matrix can dynamically adjust the mapping relationship between different granularity features,and the fused 20-joint feature $\widetilde{X_{\mathrm{out}}}$ after scale fusion can be represented as:

$$\widetilde{X_{\mathrm{out}}} = \lambda \mathrm{GCN}(A_{f_1,f_2}, f_2) + f_1 \tag{13}$$

$\mathrm{GCN}(A_{f_1,f_2}, f_2)$ represents the fused features obtained through graph convolution operation using the transformation matrix A_{f_1,f_2} on a 25-node scale. Studies have shown that the output feature maps from the shallow layers of the network can improve the quality of semantic segmentation and capture finer details [20,21]. This is because the deep feature maps of the graph convolution network often focus on high-level semantic information, while the local detail information of various skeleton parts usually exists in the shallow features, as the network goes deeper, these local details are gradually destroyed or even completely lost. Therefore, we choose to perform cross-scale feature fusion after a multi-scale graph convolution of the data, and introduce a hyperparameter λ in the fusion process to adjust the fusion ratio reasonably.

4. Experiment and Result Analysis

4.1. Experimental Dataset

The MSR Action 3D dataset is the 3D coordinates of 20 human skeleton nodes collected by the Microsoft Kinect v1 depth camera, it consists of 10 subjects performing 20 actions, each action repeated 2 to 3 times, with frame rates ranging from 10 to 100, resulting in a total of 567 action sequences. Due to the presence of highly similar action categories in this dataset, it serves as an excellent benchmark to validate the effectiveness of the algorithm proposed in this paper. The cross validation method based on subject classification is used to test the performance of the model, where subjects 1, 3, 5, 7, and 9 are used for training, and subjects 2, 4, 6, 8, and 10 are used for testing.

4.2. Experimental Environment and Settings

This experiment is implemented based on a multi-scale spatio-temporal graph convolutional network that incorporates multi-granularity features, as shown in Figure 2. The benchmark network is a stacked three-layer multi-scale spatio-temporal graph convolutional network (MS-TGCN), with input and output channels of (3, 96), (96, 192), (192, 384), and initialization representing fine-grained data initialization, CSFL is a cross scale fusion layer, GPA is a global average pooling layer, and FC is a fully connected layer. The entire model sets the batch size of the dataset to 64, and the number of iterations (epochs) for the network model is 150. The initial learning rate is 0.1, and when the number of iterations reaches 80 or 120, the learning rate decays to one tenth of the original, the weight coefficient (weight decay) is 0.0001, and the randomly discarded parameter is 0.25.

4.3. Experimental Results and Analysis

4.3.1. Comparative Experiment Using Unbiased Weighting Method

To verify the effectiveness of the proposed multi-scale adjacency matrix method, this paper designs an experiment to compare the performance differences of the model before and after the introduction of this method. The experiment uses a stacked three-layer MS-TGCN network, where MS-TGCN-D represents the multi-scale spatio-temporal graph convolution after applying the multi-scale adjacency matrix method, and the maximum value of the adjacency matrix for spatial positional relationships in the MSR Action 3D dataset is set to K=10.

As shown in Table 1, when only using the MS-TGCN network to train model parameters, the accuracy of behavior recognition roughly shows a decreasing trend with the continuous increase of K value, which well proves the bias weighting problem caused by using high-order adjacency matrices. When using the MS-TGCN-D network, the introduction of the multi-scale adjacency matrix method brings a 2.76% improvement to the network at K=6, for other values of K, it can also bring improvements ranging from 0.19% to 0.79%, thus verifies the effectiveness of introducing multi-scale adjacency matrix. However, when K=8 and K=10, the accuracy of the network decreased by 0.72% and 0.39% respectively, this is due to the highly similar characteristics of the action categories in the dataset, and the distant nodes contribute little to the recognition performance of the network, if a larger K value is adopted, the network's ability to capture the characteristics of distant nodes increases, resulting in a decrease in recognition accuracy. Therefore, in the subsequent experiments involving multi-granularity feature fusion, the value of K should not be too large, in this paper, K=6 is selected for verification in the following experiments.

Table 1. Comparison of training accuracy using multi-scale adjacency matrix method (%).

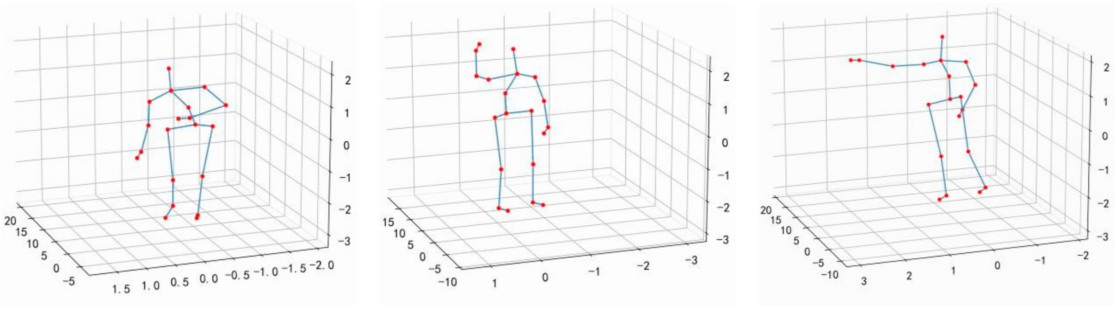
Model methods	**The value of K in K-order Adjacency Matrix**						
	K = 2	**K = 3**	**K = 4**	**K = 5**	**K = 6**	**K = 8**	**K = 10**
MS-TGCN	94.09	93.31	92.52	92.91	92.12	92.91	92.12
MS-TGCN-D	94.28	93.70	93.31	92.91	**94.88**	92.13	91.73

4.3.2. Comparative Experiments on Fusing Multi-Granularity Features

The fine-grained feature of human joint information can fully represent the refined semantic features during human movement. Therefore, this paper conducts comparative experiments on the proposed method. Refine the data of 20 human joint points in the MSR Action 3D dataset into 23 and 25-joint points respectively using the method proposed in Section 3.2, as shown in Figure 5.

The comparative experiment for fusing multi-granularity features uses MS-TGCN-D as the backbone network, fully integrating joint information of different granularities: 20-joints, 23-joints, and 25-joints; at the same time, incorporating the Cross-scale Feature Fusion Layer (CSFL) to guide the discriminative feature expression between similar behaviors learned from fine-grained data to distinguish the original granularity. The accuracy of behavior recognition using different granularities is shown in Table 2.

The accuracy rates of each behavior recognition tested by MS-TGCN-D (20-joints) and MS-TGCN-D (25-joints) are shown in Table 3. By comparison, it can be seen that fine-grained data can effectively distinguish some similar behaviors (such as drawing a fork, drawing a circle, and drawing a tick), the recognition rate of punching from the side has increased from 86.7% to 100%, and the accuracy improvement in bending action is the highest, reaching 19.5%. This is because the inserted joint points are the waist, wrist, and calves, inserting the wrist joint points allows the model to capture the differences in motion feature between drawing a fork, drawing a circle, and drawing a tick, while inserting the waist joint points helps the model capture the feature expression during bending process. However, the recognition accuracy of this model has decreased for some other behaviors (such as high waving, hand serve, and pounding). This is because these actions rely heavily on the movement state of the entire arm, and the inserted joint points make it easier for the network to capture the movement differences at the front end of the arm, resulting in a decrease in the ability to capture the motion state of the arm near the torso. The CSFL layer proposed in this paper blends different granularity features, allowing the network to fully integrate fine-grained features on the basis of its original performance, thereby improving network performance.

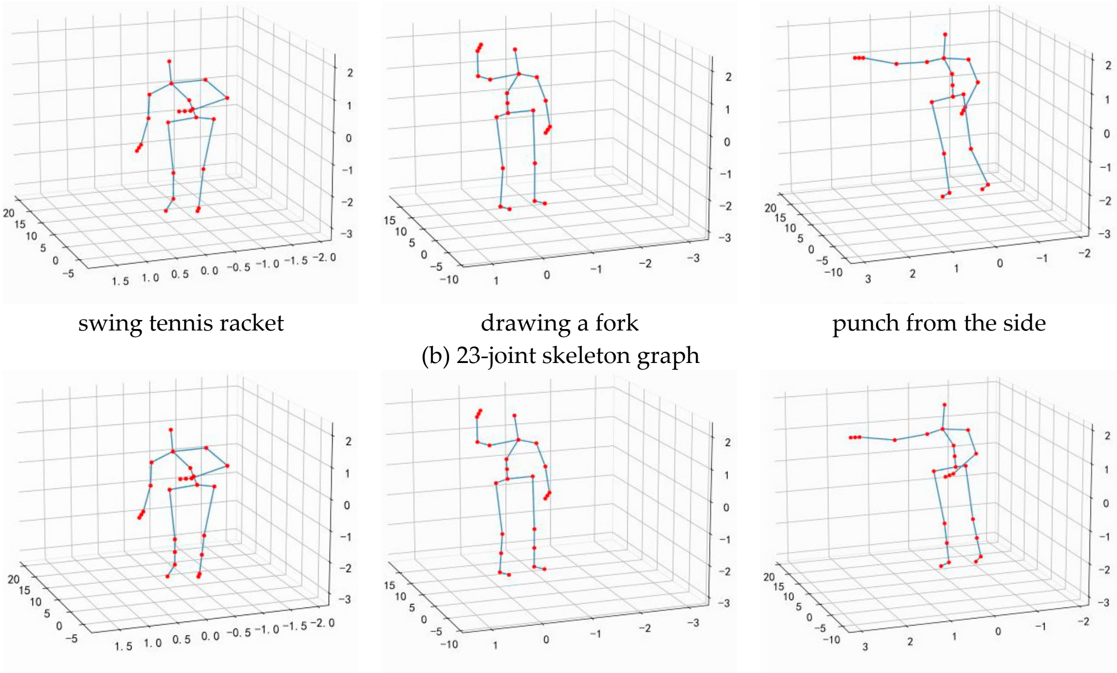


swing tennis racket

drawing a fork

punch from the side

(a) 20-joint skeleton graph



swing tennis racket drawing a fork punch from the side

(b) 23-joint skeleton graph

swing tennis racket drawing a fork punch from the side

(c) 25-joint skeleton graph

Figure 5. Skeleton graph of different granularities.

Table 2. Accuracy of recognition using different granularity data on MS-TGCN-D network.

Number of joint points / (pieces)			accuracy rate (%)
20	23	25	
✓			94.88
	✓		94.09
		✓	93.31

Table 3. Accuracy rates of each behavior recognition using data of different granularities.

MSR Action 3D behavior types	Accuracy rate of behavior recognition (%)	
	MS-TGCN-D (20-joints)	MS-TGCN-D (25-joints)
Raise your hand high(HiW)	100	81.8
Wave your hand in front of your chest(HoW)	100	100
Hammering(H)	92.3	75.0
Hand catch(HCh)	100	100
Forward punch(FP)	100	90.9
High throw(HT)	100	88.9
Drawing a fork(DX)	92.3	100
Drawing a tick(DT)	100	100
Drawing a circle(DC)	93.8	93.8
Hand Clap(HCp)	100	100
Two Hand Wave(HW)	100	100
Punch from the side(SB)	86.7	100
Bending down(B)	58.3	77.8
Kick Forward(FK)	100	100
Kick Side(SK)	100	100
Jogging(J)	100	100
Swing tennis racket(TSw)	93.8	83.3
Overhand serve(TSr)	100	93.8

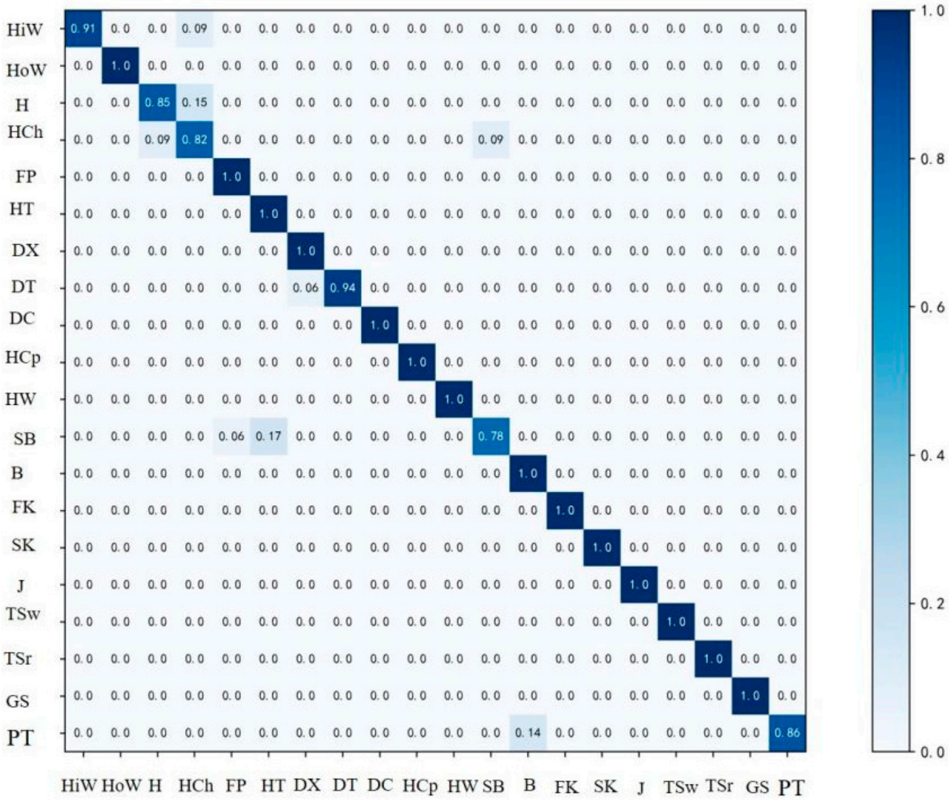
Swing a golf club(GS)	100	100
Picking up and throwing(PT)	75	75
Overall recognition accuracy rate	94.88	93.31

The cross-scale feature fusion experiment adopts three settings: fusing 20-joints with 23-joints and 25-joints respectively, and fusing all three of them simultaneously for the experiment. Among them, a Cross-scale Feature Fusion Layer (CSFL) is integrated into the backbone network (MS-TGCN-D), different fusion ratio parameters can balance the influence between coarse-grained and fine-grained. To verify the impact of fusing different granularity data under different fusion ratio parameters on network performance, this paper conducted comparative experiments on the values and fusion methods, the experimental results are shown in Table 4. According to the experimental results, it can be seen that when the network fuses three granularity data and the fusion ratio parameter is set to 0.1, the recognition accuracy reaches 95.67%, which is 0.79% higher than the accuracy without multi-granularity fusion, the network performance is optimized at this point, fully demonstrating the effectiveness of the neural network with fused multi-granularity features.

Table 4. Recognition accuracy of MS-TGCN-D (CSFL) model by fusing different numbers of joint-points under different proportional parameters.

The number of joint-points			**The value of the proportional parameter λ**	**accuracy rate (%)**
20	**23**	**25**		
✓	✓		0.1	94.09
			0.2	94.28
			0.3	93.70
✓		✓	0.1	93.31
			0.2	94.88
			0.3	92.92
✓	✓	✓	0.1	**95.67**
			0.2	92.92
			0.3	93.70

With a network recognition accuracy of 95.67%, the confusion matrix of the MSR Action 3D dataset is shown in Figure 6. As shown in the figure, the multi-scale spatio-temporal graph convolutional network that incorporates multi-granularity features can improve the discrimination rate of similar behaviors, such as drawing a fork, drawing a circle, picking up and throwing, bending down, and swinging tennis rackets on the basis of the original network, especially,the recognition accuracy of bending actions has reached 100%, which is a significant improvement compared to the original network. However, the recognition rates for behaviors such as hammering, hand catch, raise your hand high have decreased, indicating that for behaviors that rely on the entire arm movement, the arm should be endowed with more refined granularity features, rather than just focusing on the front end of the arm,this provides ideas for future work directions.



	HiW	HoW	H	HCh	FP	HT	DX	DT	DC	HCp	HW	SB	B	FK	SK	J	TSw	TSr	GS	PT
HiW	0.91	0.0	0.0	0.09	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
HoW	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
H	0.0	0.0	0.85	0.15	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
HCh	0.0	0.0	0.09	0.82	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.09	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
FP	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
HT	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DX	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DT	0.0	0.0	0.0	0.0	0.0	0.0	0.06	0.94	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
HCp	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
HW	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SB	0.0	0.0	0.0	0.0	0.06	0.17	0.0	0.0	0.0	0.0	0.0	0.78	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
FK	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
SK	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
J	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
TSw	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
TSr	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
GS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
PT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.14	0.0	0.0	0.0	0.0	0.0	0.0	0.86

Figure 6. The confusion matrix of the MSR Action 3D dataset.

4.3.4. Comparison Experiment with other Models

In order to better verify the improvement of the model on behavior recognition performance, this paper compared and analyzed the recognition accuracy with existing behavior recognition methods on the MSR Action 3D dataset. The comparison results are shown in Table 5.

The multi-scale spatio-temporal graph convolutional network proposed in this paper, which integrates multi-granularity features, achieves a behavior recognition accuracy of 95.67% on the MSR Action 3D dataset, and its experimental results are superior to most existing behavior recognition methods. Compared with the methods proposed in references [20,22], the accuracy has been improved by 2.04% and 3.77% respectively; compared with the adaptive skeleton center point method proposed in reference [21], the accuracy has been improved by 7.2%; compared with the method of combining graph convolution with Long Short-Term Memory (LSTM) networks[10] and the multi-view depth motion map method STACOG [23], the accuracy is improved by 1.17% and 2.27% respectively; compared with the enhanced data-driven algorithm proposed in reference [24] and the method of using point cloud data as input for behavior recognition [25], the accuracy has been improved by 0.86% and 0.49% respectively; compared with the fusion multi-modal data feature method proposed in reference [26], the accuracy has been improved by 3.76%. By comparison, it can be seen that the algorithm proposed in this paper has a high recognition accuracy in using 3D human skeleton information for human behavior recognition, and also has strong competitiveness compared to existing methods.

Table 5. Comparison of recognition accuracy with other methods on the MSR-Action 3D dataset.

method	accuracy rate (%)
Yang et al. [20]	93.63
adaptive skeleton center point [21]	88.47
Agahian et al. [22]	91.90
Zhao et al. [10]	94.50

STACOG [23]	93.40
Zhang et al. [24]	94.81
Wu et al. [25]	95.18
You et al. [26]	91.91
Ours	**95.67**

5. Conclusion

The fine-grained features of human skeleton data can represent semantic features of different levels during behavioral process, and integrating the motion features at different granularity levels can effectively improve the recognition effect of the network for similar behaviors. This paper proposes a multi-scale spatio-temporal graph convolution method that integrates multi granularity features for human behavior recognition. The skeleton fine-grained partitioning strategy initializes human skeleton data into data streams of different granularities, and the spatio-temporal graph convolutional network with multi-scale adjacency matrices can effectively improve the network's spatio-temporal representation capabilities. The adaptive cross-scale fusion layer guides the model to learn discriminative feature expressions between similar behaviors with fine-grained features, thereby improving the robustness of the network in recognize similar behaviors. The experimental results on the MSR Action 3D dataset show that the accuracy of our algorithm for behavior recognition is superior to most existing methods, thus verifying the effectiveness of our algorithm.

At present, most models extract features from the global perspective, and their ability to capture local differences between fine-grained actions is insufficient, the focus of further research is on how to extract representations of locally sensitive actions, so as to better characterize the small local differences between fine-grained actions. In the future, the application of the algorithm proposed in this paper will be studied under different scene features, considering the characteristics of human motion processes in different scenes and using different features of coarse and fine granularity reasonably to improve the accuracy of skeleton behavior recognition.

Author Contributions: Conceptualization, Y.W., T.S.; methodology, Y.Y.; investigation and validation, Y.W..; writing—original draft preparation, Z.H.; writing—review and editing, T.S. and Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was supported by the following projects: Basic and Frontier Research Program of Chongqing Science and Technology Bureau (Grant No.CSTC2021JCYJ-MSXMX0348), Science and Technology Research of Chongqing Municipal Education Commission(Grant No.KJQN202305804),The first batch of young backbone teacher training plan for Chongqing Public Transport Vocational College..

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. P . M. Pilarski, A. Butcher, M. Johanson, M. M. Botvinick, A. Bolt, and A. S. R. Parker, ‘‘Learned human-agent decision-making, communication and joint action in a virtual reality environment,’’ 2019, arXiv:1905.02691.
2. Shi L, Zhou Y, Wang J, et al. Compact global association based adaptive routing framework for personnel behavior understanding[J]. Future Generation Computer Systems, 2023, 141:514-525.
3. Sudha M R, Sriraghav K, Jacob S G, et al. Approaches and applications of virtual reality and gesture recognition: A review[J]. International Journal of Ambient Computing and Intelligence (IJACI), 2017,8(4): 1-18.
4. Duan H, Zhao Y, Chen K, et al. Revisiting skeleton-based action recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:2969-2978.
5. Liu Congcong, Ying Jie, Yang Haima, et al. Improved human action recognition approach based on two-stream convolutional neural network model [J]. The Visual Computer, 2021, 37, 1327-1341.

6. Duan H, Zhao Y, Chen K. Revisiting Skeleton-Based Action Recognition[C]. Proceedings of the IEEE/CVF ConferenceonComputerVisionandPatternRecognition,2022: 2969-2978
7. Liu J, Shahroudy A, Xu D, et al. Spatio-temporal LSTM with trust gates for 3D human action recognition[C] //Proceedings of the 14th European Conference on Computer Vision. Heidelberg: Springer, 2016: 816-833
8. Wei S H, Song Y H, Zhang Y L. Human skeleton tree recurrent neural network with joint relative motion feature for skeleton based action recognition[C] //Proceedings of the IEEE International Conference on Image Processing (ICIP). Los Alamitos: IEEE Computer Society Press, 2017: 91-95
9. Zheng W, Li L, Zhang Z X, et al. Relational network for skeleton-based action recognition[C] //Proceedings of the IEEE International Conference on Multimedia and Expo(ICME). Los Alamitos: IEEE Computer Society Press, 2019:826-831
10. Zhao, R.; Wang, K.; Su, H.; Ji, Q. Bayesian graph convolution lstm for skeleton-based action recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019.
11. Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]//Thirty-second AAAI conference on artificial intelligence. 2018.
12. Shi L, Zhang Y, Cheng J, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 12026-12035.
13. Li M S, Chen S H, Chen X, et al. Actional-structural graph convolutional networks for skeleton-based action recognition[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Los Alamitos: IEEE Computer Society Press, 2019: 3590-3598
14. Chaolong Li, Zhen Cui, Wenming Zheng, Chunyan Xu, and Jian Yang. Spatio-temporal graph convolution for skeleton based action recognition. In Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
15. Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3595–3603, 2019.
16. Renjie Liao, Zhizhen Zhao, Raquel Urtasun, and Richard SZemel. Lanczosnet: Multi-scale deep graph convolutional networks. arXiv preprint arXiv:1901.01484, 2019.
17. Cheng K, Zhang Y, He X, et al. Skeleton-based action recognition with shift graph convolutional network[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020:183-192.
18. Li W, Liu X, Liu Z, et al. Skeleton-based action recognition using multi-scale and multi-stream improved graph convolutional network[J]. IEEE Access, 2020, 8: 144529-144542.
19. Liu Z, Zhang H, Chen Z, et al. Disentangling and unifying graph convolutions for skeleton-based action recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Piscataway, NJ: IEEE, 2020: 143-152.
20. Yang, Y .; Deng, C.; Gao, S.; Liu, W.; Tao, D.; Gao, X. Discriminative multi-instance multitasks learning for 3D action recogni-tion. IEEE T rans. Multimed. 2017, 19, 519–529.
21. Ran Xianyu, Liu Kai, Li Guang, et al. Human action recognition algorithm based on adaptive skeleton center[J]. Journal of Image and Graphics, 2018,23(04): 519-525.
22. Agahian, S.; Negin, F.; Köse, C. Improving bag-of-poses with semi-temporal pose descriptors for skeleton-based action recognition. Vis. Comput. 2019, 35, 591–607.
23. Bulbul M F, Tabussum S, Ali H, et al. Exploring 3D human action recognition using STACOG on multi-view depth motion graphs sequences[J]. Sensors, 2021, 21(11): 3642-3651.
24. Zhang, C.; Liang, J.; Li, X.; Xia, Y .; Di, L.; Hou, Z.; Huan, Z. Human action recognition based on enhanced data guidance and key node spatial temporal graph convolution. Multimed. T ools Appl. 2022, 81, 8349–8366.
25. Wu, Q.; Huang, Q.; Li, X. Multimodal human action recognition based on spatio-temporal action representation recognition model. Multimed. T ools Appl. 2022, 81, 1–22.
26. You, K., Hou, Z., Liang, J., Lin, E., Shi, H., & Zhong, Z. (2024). A 4D strong spatio-temporal feature learning network for behavior recognition of point cloud sequences. Multimedia Tools and Applications. https://doi.org/10.1007/s11042-023-18045-3

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.