Decoupled 2S-AGCN Human Behavior Recognition Based on New Partition Strategy

Liu Qiuming[1,2], Chen Longping[1(✉)], Wang Da[1], Xiao He[1,2], Zhou Yang[3], and Wu Dong[3]

[1] School of Software Engineering, Jiangxi University of Science and Technology, Nanchang 330013, China
liuqiuming@jxust.edu.cn, {6720220666, 6720220658}@mail.jxust.edu.cn

[2] Nanchang Key Laboratory of Virtual Digital Factory and Cultural Communications, Nanchang 330013, China

[3] Information and Communication Branch of Jiangxi Electric Power Co., Ltd., Nanchang 330095, China

Abstract. Human skeleton point data has better environmental adaptability and motion expression ability than RGB video data. Therefore, the action recognition algorithm based on skeletal point data has received more and more attention and research. In recent years, skeletal point action recognition models based on graph convolutional networks (GCN) have demonstrated outstanding performance. However, most GCN-based skeletal action recognition models use three stable spatial configuration partitions, and manually set the connection relationship between each skeletal joint point. Resulting in an inability to better adapt to varying characteristics of different actions. And all channels of the input X features use the same graph convolution kernel, resulting in coupling aggregation. Contrary to the above problems, this paper proposes a new division strategy, which can better extract the feature information of neighbor nodes of nodes in the skeleton graph and adaptively obtain the connection relationship of joint nodes. And introduce Decoupled Graph Convolution (DC-GCN) to each partition to solve the coupled aggregation problem. Experiments on the NTU-RGB+D dataset show that the proposed method can achieve higher action recognition accuracy than most current methods.

Keywords: 2S-AGCN · New partition strategy · DC-GCN · Action recognition · NTU RGB+D

1 Introduction

As a research hotspot in the field of computer vision, human action recognition has a wide range of applications in many fields such as video surveillance, human-computer interaction, and social security. Therefore, people are more and more interested in research in this field. According to the form of input data, action recognition methods can be

C. Wu et al. (Eds.): MONAMI 2023, LNICST 559, pp. 70–82, 2024.
https://doi.org/10.1007/978-3-031-55471-1_6

roughly divided into two categories: one is image-based, and the other is skeleton-based. In image-based recognition methods, RGB data is usually used as input to realize action recognition by extracting image features. In the recognition method based on human skeleton data, the skeleton data composed of two-dimensional or three-dimensional coordinates of joint points of the human body is extracted, and the action is recognized by feature extraction technology [1].

Human action recognition is a complex task because it requires a deep understanding of image content. Due to the diversity and complexity of human behavior and postures, as well as possible occlusions and other problems. Human action recognition is more challenging and complex than merely recognizing or detecting objects in images. The study of image information works well under simple background conditions. However, in reality, it may be affected by noise such as illumination. The action recognition method based on human skeleton information uses posture information to represent human characteristics, which can effectively reduce the impact of illumination [2]. With the rise and wide application of depth cameras, it becomes easier to obtain precise coordinates of joint points of the human skeleton. The action recognition method based on skeleton data has the characteristics of robustness and insensitivity to changes in lighting conditions, and can show better and excellent performance when given accurate joint point coordinates.

Researchers mainly use three deep techniques to learn actions in skeleton sequences, namely traditional convolutional neural network (CNN), recurrent neural network (RNN), and graph convolutional network (GCN). CNN has a strong ability to extract spatio-temporal features, while RNN is suitable for modeling temporal information [3]. Both approaches, RNN and CNN, have been widely used in skeleton-based action recognition and achieved impressive results. However, these two types of methods always represent bones in a grid-like manner, which cannot fully express the spatial structure information between human joints. Recently, graph convolutional networks (GCNs) have gradually received more attention, which are very suitable for processing non-Euclidean data. Skeletons can be naturally represented as graphs in non-Euclidean spaces. GCN-based methods have made substantial improvements in skeleton-based action recognition tasks. Therefore, Yan et al. first introduced the graph convolutional network into skeletal action recognition, and proposed the spatio-temporal graph convolutional network (ST-GCN) [4], which uses the natural connection relationship between human joints for action modeling, which does not require Manually design and divide skeleton parts or make human skeleton joint point traversal rules, so this method achieves better performance than previous methods.

In recent years, GCN has been rapidly developed and applied to process graph data. It usually has two kinds of construction ideas, which follow different principles. One is the spectral domain idea, whose principle is to perform graph convolution similarly in the frequency domain with the help of Fourier transform; the other is the spatial domain idea, whose principle is to use convolution directly on the nodes and their neighborhoods of the topological graph filter to extract features [5]. The graph convolutional network (GCN) that has emerged in recent years can make full use of the connection relationship between nodes to model data, which is very suitable for action recognition applications based on skeleton point data. Shi et al. [6] proposed the 2S-AGCN network, which

adds an adaptive topological graph to each graph convolutional layer to enhance the long-range spatial modeling capability of the graph convolutional layer. Zhang et al. [7] proposed the SGN network, which uses the semantic information of human body joints and frames to enrich the expressive ability of skeleton features, thereby improving the recognition accuracy of the model. In any case, neither RNN network nor CNN network can fully characterize the spatial structure of skeleton data, because skeleton data is not a sequence of vectors or a two-dimensional grid, which has the structure of a graph of natural connections of human body structure. Compared with the former two methods, the GCN-based method does not need to manually divide the skeleton into multiple parts and design joint traversal rules, and can preserve the skeleton topology in the process of modeling skeleton space and time dependencies. Therefore, the GCN-based action recognition method has more advantages in modeling the spatiotemporal characteristics of the skeleton and has gradually become the preferred framework in this field.

However, most GCN-based methods still have some shortcomings, which have not been considered by current research. The above GCN for skeleton-based action recognition models all use stable three spatial configuration partitions and manually set the connection relationship between each skeletal joint point, which cannot better adapt to the changing characteristics of different actions. To solve this problem, in order to better extract the feature information of neighbor nodes of the nodes in the skeleton graph and adaptively obtain the connection relationship of joint nodes, this paper proposes a new partition strategy based on the original 2S-AGCN space-time model. In addition, for the coupling aggregation phenomenon caused by the feature channels of the input X sharing the same adjacency matrix, by introducing decoupling graph convolution to each partition, the feature channel information is fully utilized to improve the network's ability to model human bone information.

2 Related Work

2.1 ST-GCN Model

The input of the spatiotemporal graph convolutional network can be expressed as a spatiotemporal graph G = (V, E), as shown on the left side of Fig. 1. $V = \{V_{ti} \mid t = 1, \cdots, T; i = 1, \cdots, v\}$, T represents the sequence frame, v represents the joint position, and V_{ti} represents the 2D or 3D coordinate data of the i-th joint in the t-th frame [8]. E includes E_S and E_F, $E_S = \{V_{ti}V_{tj} \mid (i, j) \in H\}$ represents the connection between human joints in a skeleton data frame, H is the joint natural connection set, $E_F = \{V_{ti}V_{(t+1)i}\}$ means that the same joints are interconnected in the time latitude. After ST-GCN obtains the skeleton sequence data composed of coordinates, it models the structured information between these joints along the space and time dimensions. The space dimension refers to the dimension where the joints in the same frame are located, and the time dimension refers to a certain One row is the dimension in which the same joint is located for all frames.The spatiotemporal graph convolutional neural network is composed of 9 ST-GCN units, each ST-GCN unit is composed of GCN and temporal convolutional network (temporal convolutional network, TCN), and a residual mechanism is added between ST-GCN units (residual).The main process is: given a skeleton sequence of an action video, first construct the graph structure data

expressing the sequence, and use it as the input of ST-GCN; then extract high-level spatiotemporal features through a series of spatiotemporal graph convolution operations; Finally use the Softmax classifier gets the classification result.The core idea of ST-GCN is to combine the graph convolutional network (Graph Convolutional Network, GCN) with the temporal convolutional network (Temporal Convolutional Network, TCN), among them, GCN convolution space dimension data, TCN convolution time dimension The data.The skeleton spatiotemporal graph is shown in Fig. 1.

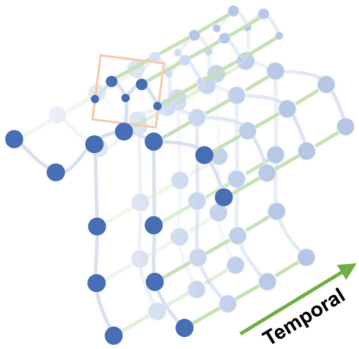


Fig. 1. Skeleton spatiotemporal diagram

The convolution of the single-frame skeleton graph above is represented by the following for Eq. (1):

$$f_{out} = \sum_{k}^{Kv} W_k (f_{in} A_k) \odot M_k \tag{1}$$

Among them, f_{in} is used as the input of the network, and the input is the skeleton sequence V in the space-time graph, and f_{out} is the output of the ST-GCN network, which is the behavior category output by the fully connected layer. K_v is the number of subsets divided by neighbor nodes, that is, the number of partitions. The partition strategy used by ST-GCN is space configuration partition, and the size of K_v is set to 3. The three configuration partitions are respectively represented by the adjacency matrix A_k, where $k \in \{1, 2, 3\}$, and the size of A_k is $N \times N$. W_k is the weight function of graph convolution, which is a two-dimensional convolution with a convolution kernel size of 1×1. M_k is an $N \times N$ attention matrix used to learn the importance weights of node connections in the adjacency matrix, and $\odot$ represents the dot product.

ST-GCN proposes three partition strategies, namely Uni-labelling, Distance partitioning, and Spatial configuration partitioning. Uni-labelling is the simplest strategy, where each node has the same label (0) and $K = 1$; Distance partitioning is also simple, it divides the adjacent nodes according to their distance from the root node, if $D = 1$, then this means that the root node is always 0 and all other nodes are 1. This enables the algorithm to simulate local differential behavior. $K = 2$ and lti(vti) = d(vti, vti); Spatial configuration partitioning division is a little more complicated, and the 1 neighborhood of node i is divided into three subsets. The first subset is the node i itself, the second

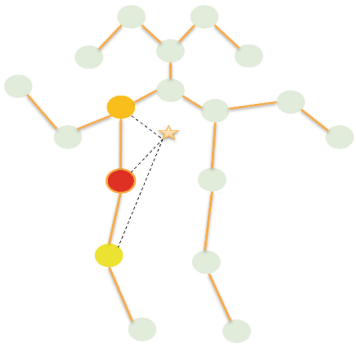


Fig. 2. Spatial configuration partitioning

subset is a set of neighbor nodes, which is closer to the center of the skeleton than node i, and the third subset is the set of neighbor nodes farther away from the center of the skeleton than node i, respectively denoting Motion characteristics such as static state, centripetal motion and centrifugal motion. K represents the number of partitions, and the space-time graph node neighborhood span is represented by D. The strategy adopted by ST-GCN is Spatial configuration partitioning, as shown in Fig. 2. This partition strategy can be expressed by Eq. 2 below.

$$l_{ti}(v_{ti}) = \begin{cases} 0 \ if \ r_j = r_i \\ 1 \ if \ r_j < r_i \\ 2 \ if \ r_j > r_i \end{cases} \tag{2}$$

where r_j is the distance from adjacent node j to the root node, and r_i is the distance from root node i to the center of gravity.

2.2 2S-AGCN Model

Since the adjacency matrix A_k in ST-GCN is shared in each spatiotemporal graph convolutional network layer, and only the natural connection relationship of the human body can be used, non-existent connections cannot be established. However, many bone nodes that are not directly connected in the process of action will also have actions related to each other. For example, clapping and other actions, the bone joint points between the two hands will have a relationship, that is, there will be a connection relationship other than the natural structure joint points of the human body. At this time, the predefined adjacency matrix that only includes the natural connection relationship of human bone points will not guarantee learning. The obtained bone point connection relationship is optimal. In order to solve the problem that ST-GCN only contains first-order information of bones, lacks second-order information and uses a stable graph structure, Shi et al. [14] proposed an adaptive graph volume called 2S-AGCN Productive dual-stream network structure. 2S-AGCN still follows the method of constructing spatio-temporal graph in ST-GCN, which is used to extract the features of joints in spatial and temporal dimensions. However, different from ST-GCN, Adaptive Spatio-Temporal Graph

Convolutional Network (2S-AGCN) tries to adaptively learn non-existing connection relations while trying to learn data correlation between samples.The form of Eq. (1) is changed so that the structure of the graph can be adjusted adaptively.

$$f_{out} = \sum_{k}^{Kv} W_k f_{in}(A_k + B_k + C_k) \tag{3}$$

In Eq. (3), A_k, which is the same as in Eq. (1), represents the normalized adjacency matrix of the physical structure of the human body. During training, elements of B_k and other parameters are parameterized and optimized together. It does not have any effect on the value of B_k, which indicates that the graph is learned entirely from the training data. With this data-driven approach, the model can fully learn the graph to complete the recognition task and be more personalized to the different information contained in different layers. It can be noted that the elements in the matrix can have arbitrary values. It not only indicates whether two joints are connected, but also how strong the connection is. C_k is the attention map between sample data, which can be obtained by Eq. (4), making the model fully data-driven, where $W\theta$ and $W\varphi$ are the weights of the Gaussian embedding function. The output of the two Gaussian embedding functions is multiplied to obtain a joint similarity weight matrix with a size of $N \times N$. The similarity matrix is passed through the softmax activation function [9] to obtain the similarity score between any two joints in the sample. Different samples will have considerable differences even with the same motion characteristics under the influence of subjects and cameras.

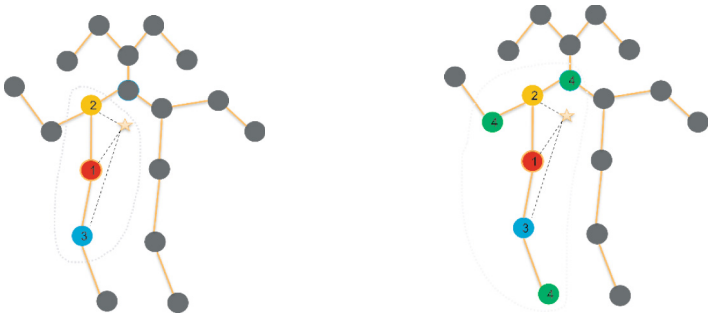
$$C_k = \tanh(f_{\text{in}}^T W_{\theta k}^T W_{\varphi k} f_{\text{in}}) \tag{4}$$

3 Decoupled 2S-AGCN Based on New Partition Strategy

3.1 2S-AGCN with New Partition Strategy

Most of the current GCN-based models often use a stable space configuration partition strategy and manually set the connection relationship between each bone joint point, which cannot better adapt to the changing characteristics of different actions. The 2S-AGCN also uses the same spatial configuration partitioning as the above-mentioned ST-GCN. To solve this problem, under the condition that the time domain is consistent with the traditional strategy, the new partition strategy proposed in this paper expands the node neighborhood span D in the space domain to 2, and changes the partition from the original three partitions to four partitions. Different partition strategies are equivalent to changing the size of the convolution kernel. At the same time, the proposed segmentation strategy can cover most of the motion joint points in the human skeleton, such as arms, thighs, etc., so it is no longer limited to adjacent joint points, but can be effectively extended to other joints [10]. The original partition strategy and the new partition strategy are shown in Fig. 3:

The division strategy of the original space configuration is shown in Fig. 3(a). Under the original partition strategy, the space-time graph is divided into three parts according to the distance from each node to the center of gravity of the skeleton: root node, centripetal



(a) Original partition strategy (b) New partition strategy

Fig. 3. Comparison of partition strategies

node, and centrifugal node [11]. 1 is the root node, 2 is the centripetal node, and 3 is the centrifugal node. The center of gravity of the skeleton is represented by a star in the figure, which is the average coordinate of all joints in the coordinate system. The new partitioning strategy proposed in this paper is shown in Fig. 3(b). Expressed as follows:

1: Root node.
2: Centripetal node, closer to the center of gravity of the skeleton than the root node.
3: Centrifugal node, which is further away from the center of gravity of the skeleton than the root node.
4: A node whose root node neighborhood span D is 2.

The implementation process of the adaptive graph convolution layer in 2S-AGCN is shown in Fig. 4 where each layer has three graph subsets A_k, B_k, and C_k. The orange box indicates that the parameters can be learned, and k_v indicates the number of partitions. In 2S-AGCN, k_v is 3. After using the new partition strategy, the k_v of the adaptive graph convolution layer changes from 3 to 4.

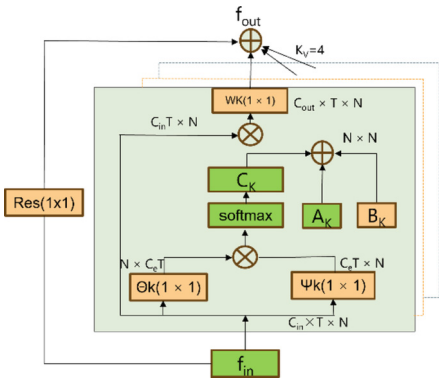


Fig. 4. New Adaptive Convolutional Layers

3.2 Decoupling GCN

Graph convolution consists of two matrix multiplication processes: AX and XW. AX calculates aggregated information between different skeletons, so we call it spatial aggregation. XW calculates the correlation information between different channels, so we call it channel correlation.

As the picture shows. In Fig. 5, the spatial aggregation (AX) can be decomposed to compute the aggregation on each channel separately. Note that all channels of feature X share an adjacency matrix A (drawn in the same color), which means that all channels share the same convolution kernel. We call this coupled aggregation. Since the human body has multiple degrees of freedom, the correlation between joints is very complex, and the correlation between different actions is also different, which limits the expressive ability of graph convolution spatial aggregation. While most of the existing GCN-based skeletal action recognition methods use coupling aggregation, such as ST-GCN [4], non-local adaptive GCN, AS-GCN [12], Directed-GNN [13], 2S-AGCN [6]. We collectively refer to them as coupled graph convolutions.

Fig. 5. Coupling aggregation in GCNS

Fig. 6. Decoupling aggregation in GCNS

In this paper, we adopt Decoupled Graph Convolution (DC-GCN) for skeletal action recognition, where different channels have independent trainable adjacency matrices, as shown in Fig. 6. Decoupled graph convolutions greatly increase the diversity of adjacency matrices. Similar to the redundancy of CNN kernels [14], decoupled graph convolutions may introduce redundant adjacent matrices. Therefore, we divide the channels into g groups. Channels in a group share a trainable adjacency matrix. When $g = C$, each channel has its own spatial aggregation kernel, which leads to a large number of redundant parameters; when $g = 1$, the decoupled graph convolution degenerates into a coupled graph convolution [15]. So we can temporarily set g to 8. The equation for decoupling graph convolution is as follows:

$$\mathbf{X}^{'} = \overset{\sim}{\mathbf{A}}^{d}_{:,:,1}\mathbf{X}^{\mathbf{w}}_{:,\lfloor \frac{\mathbf{C}}{\mathbf{g}} \rfloor} \| \overset{\sim}{\mathbf{A}}^{d}_{:,:,2}\mathbf{X}^{\mathbf{w}}_{:,\lfloor \frac{\mathbf{C}}{\mathbf{g}} \rfloor:\lfloor \frac{2\mathbf{C}}{\mathbf{g}} \rfloor} \| \cdots \| \overset{\sim}{\mathbf{A}}^{d}_{:,:,9}\mathbf{X}^{\mathbf{w}}_{:,\lfloor \frac{(\mathbf{g}-1)\mathbf{C}}{\mathbf{g}} \rfloor:} \tag{5}$$

where $X^w = XW$, $A^d \in R^{n \times n \times g}$ is the decoupled adjacency matrix. The indices of A^d and X^w are in Python notation, representing channel-level connections.

DC-GCN can be naturally extended to the case of multiple partitions by introducing a decoupled graph convolution to each partition. Note that our DC-GCN differs from the multi-partition strategy, which integrates multiple graph convolutions with different adjacency matrices. Combining DC-GCN with the new partitioning strategy proposed above for experiments shows the complementarity between multi-partitioning strategies and DC-GCN.

4 Experiment and Analysis

This paper uses the large-scale public behavior dataset NTU-RGB+D [16] to conduct ablation experiments to verify the effectiveness of the model components; compare the method with some current mainstream and advanced methods to verify the performance level of the method proposed in this paper. Experiments show that the model has achieved high recognition accuracy on the NTU-RGB+D dataset, which verifies the effectiveness of the method proposed in this paper. The experimental platform used is: Window10 system, the CPU is i5-12500H, the graphics card is RTX3090, the memory is 8 GB, and the deep learning framework is Pytorch.

All models in this paper use stochastic gradient descent (SGD) optimizer for training, and set the momentum is 0.9, and the weight decay is 0.0001. The training epochs is set to 50. The initial learning rate is set to 0.1, and the learning rate decays with a coefficient of 0.1 at the 20th epoch and 40th epoch, and the batch size is set to 32.

4.1 Datasets

NTU-RGB+D. NTU RGB+D contains 56,880 motion clips grouped into 60 categories. They invited 40 different volunteers to perform the moves. Use three cameras simultaneously to capture three different horizontal views of the same action. These actions were recorded by 40 volunteers in a laboratory environment using three cameras simultaneously. Detect and provide annotations through Kinect depth sensors, and provide 3D joint positions (X, Y, Z) in the camera coordinate system, mainly including 25 joint points of the human body.The 3D coordinates of 25 body joints are captured to form a skeleton sequence. The structure diagram of NTU human body is shown in Fig. 7.The authors of this dataset used two evaluation schemes:

(1) cross-subject (CS): The training set contains 40,320 samples from 20 subjects, and the test set has 16,560 samples from other subjects.
(2) cross-view (CV): The training set contains 37,920 samples from two camera views, while the test set from another camera view contains 18,960 samples.

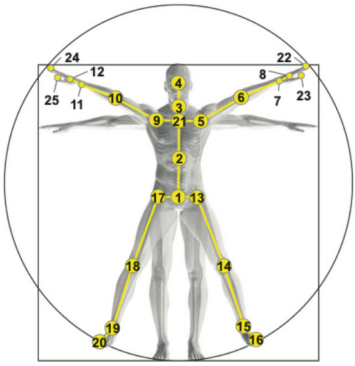


Fig. 7. NTU Dataset Anatomy of the Human Body

4.2 Ablation Experiment

In order to verify the improvement effect of the different improvements proposed in this paper on the original 2S-AGCN model, comparative experiments were carried out on the NTU-RGB + D dataset using new partition strategies and using decoupled graph convolution. The specific experimental results are shown in Table 1, where X-Sub and X-View represent the results obtained from different targets or different camera angles for the test samples, respectively. In order to verify whether the new partition strategy can improve the performance of the model, in the original 2S-AGCN model, the original partition strategy is replaced with the new partition strategy, namely 2S-AGCN + NPS in the table. At the same time, in order to test the performance of decoupled graph convolution (DC-GCN), DC-GCN is added to the original 2S-AGCN, that is, 2S-AGCN + DC in the table. Our proposed model is represented by 2S-AGCN + NPS + DC.

Table 1. Accuracy comparison experiment results of different model structures

Backbone	+ NPS	+ DC	X-Sub (%)	X-View (%)
2S-AGCN			88.5%	95.1%
2S-AGCN	✔		88.7%	95.2%
2S-AGCN		✔	88.8%	95.3%
2S-AGCN	✔	✔	**90.1%**	**95.8%**

It can be clearly seen from Table 1 that the accuracy rate on the NTU-RGB + D data set has been improved after using the new division strategy, which has increased by 0.2% in the X-Sub evaluation method and 0.1% in the X-View evaluation method. At the same time, it can also be seen that the accuracy rates on X-Sub and X-View have increased by 0.3% and 0.2% respectively after using DC-GCN.

Under the CV evaluation benchmark of the NTU-RGB + D dataset, the confusion matrix of the validation set is shown in Fig. 8. There are a total of 60 behaviors in the

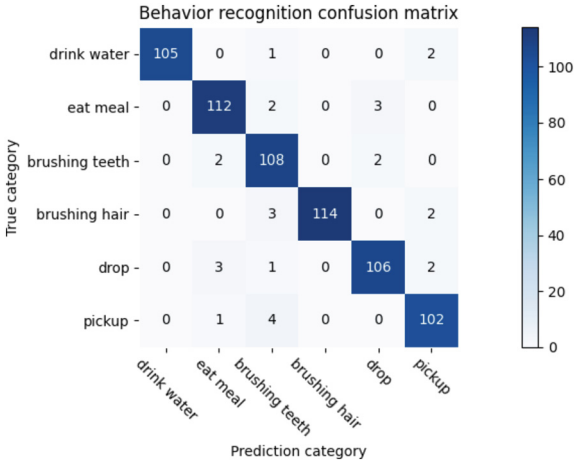


Fig. 8. Behavior recognition confusion matrix

data set. For the convenience of display, the recognition results of the first 6 behaviors are drawn as a confusion matrix for analysis. The recognition accuracy of these six behaviors is above 95%, confirming the effectiveness of our proposed two modules.

4.3 Two-Stream Fusion Experiment

In addition to the first-order information (coordinates, confidence), bone information, its second-order information (length and direction of the bone) is also very important [17]. We use Joint to extract the Bone information of the skeleton, and train the Joint flow and Bone flow information respectively. This is the meaning of 2S in 2S-AGCN. Finally, the results of the two-stream recognition are fused to obtain the recognition result of the final model. In the X-View evaluation mode of the NTU-RGB + D dataset, multiple experiments have shown that bone data can improve the performance of the method. As shown in Table 2.

Table 2. Dual-stream network performance under X-View

Modal	Accuracy(%)
J stream	94.72%
B stream	94.34%
2 stream	95.8%

4.4 Compared with Other Behavior Recognition Methods

In order to verify the performance level of the method proposed in this paper, we compared the recognition accuracy of the behavior recognition method proposed in this

paper with several current research popular recognition technologies on the NTU-RGB + D dataset. The methods involved in the comparison include CNN-based methods and GCN-based methods. Table 3 below lists the recognition accuracy of these algorithms on the NTU-RGB + D dataset. It can be seen from Table 3 that the accuracy of 2S-AGCN under the new partition strategy and decoupled graph convolution is significantly higher than other behavior recognition methods. Compared with other methods based on CNNs or RNNs, the recognition accuracy of this method has been greatly improved.In the NTU-RGB + D dataset X-Sub division mode, the recognition rate is 0.6 percentage points higher than that of the 2S-AGCN model, and in its X-View division mode, the recognition rate is 0.7 percentage points higher than that of the 2s-AGCN model. This shows that the decoupled 2S-AGCN based on the new partition strategy proposed in this paper can better improve the performance of the model on large data sets.

Table 3. Comparison of recognition accuracy between the proposed method and other methods

Methods	X-Sub (%)	X-View (%)
Deep LSTM	60.7%	67.3%
Two-Stream 3DCNN	66.8%	72.6%
ST-LSTM	69.2%	77.7%
TCN	74.3%	83.1%
ST-GCN	81.5%	88.3%
2S-AGCN	88.5%	95.1%
Ours	**90.1%**	**95.8%**

5 Conclusion

The traditional graph convolutional network only uses three stable spatial configuration partitions and manually sets the connection relationship between the joint points of the bones, which cannot better adapt to the changing characteristics of different actions and ignores the information of non-adjacent nodes. A new division strategy is proposed, which can better extract the feature information of the neighbor nodes of the nodes in the skeleton graph and adaptively obtain the joint connection relationship. In addition, for the coupling aggregation phenomenon caused by using the same graph convolution kernel for all channels of the input X feature, by introducing decoupled graph convolution to each partition, the channel information of the feature is fully utilized to improve the network's modeling of human bone information. Ability. In order to verify the effectiveness of the method, extensive experiments are conducted on the NTU-RGB + D dataset. The results show that the method proposed in this paper can obtain higher action recognition accuracy than most current literatures, and the accuracy rates of 90.1% and 95.8% were respectively achieved under the X-Sub and X-View division methods of the NTU-RGB + D dataset. The next step of research will be to introduce the attention module to improve the ability of the network to model spatio-temporal features.

References

1. Xuanye, L., Xingwei, H., Jingong, J., et al.: Human action recognition method combining multi-attention mechanism and spatiotemporal graph convolutional network. J. Comput. Aided Des. Graph. **33**(07), 1055–1063 (2021)
2. Hualei, X., Yingqiang, D., Meng, G., et al.: Skeletal action recognition based on multi-partition spatiotemporal graph convolutional network. Signal Process. **38**(02), 241–249 (2022). https://doi.org/10.16798/j.issn.1003-0530.2022.02.003
3. 0036 L J,Shahroudy A,Xu D, et al. Spatio-Temporal LSTM with Trust Gates for 3D Human Action Recognition.[J]. CoRR,2016,abs/1607.07043
4. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-2018) and the Thirtieth Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-2018) (2018)
5. Haoli, Z.: Action recognition based on spatial-temporal graph convolutional network with fusion of geometric features. Computer Syst. Appl. **31**(10), 261–269 (2022)
6. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
7. Zhang, P., Lan, C., Zeng, W., Xing, J., Xue, J., Zheng, N.: Semantics-guided neural networks for efficient skeleton-based human action recognition. In:2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
8. Chen, W., Qiang, S., Hongyu, N., et al.: Abnormal behavior recognition based on skeleton sequence extraction. Comput. Syst. Appl. **31**(11),215–222 (2022).https://doi.org/10.15888/j.cnki.csa.008773
9. Mikolov T,0010 C K,Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. CoRR,2013,abs/1301.3781
10. Wang, Q., Zhang, K., Asghar, M.A.: Skeleton-based ST-GCN for human action recognition with extended skeleton graph and partitioning strategy. IEEE Access **10**, 41403–41410 (2022)
11. Wu, J., Wang, L., Chong, G., Feng, H.: 2S-AGCN human behavior recognition based on new partition strategy. In: 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) (2022)
12. Li, Q., Han, Z., Wu, X.M.: Deeper insights into graph convolutional networks for semi-supervised learning. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
13. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Skeleton-based action recognition with directed graph neural networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
14. Molchanov, P., Tyree, S., Karras, T., Aila, T., Kautz, J.: Pruning convolutional neural networks for resource efficient inference. arXiv preprint arXiv:1611.06440(2016)
15. Cheng, K., Zhang, Y., Cao, C., Shi, L., Cheng, J., Lu, H.:GCN with dropgraph module for skeleton-based action recognition. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, JM. (eds.) Computer Vision – ECCV 2020. ECCV 2020. LNCS, vol. 12369, 536–553. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58586-0_32
16. Fernando, B., Gavves, E., José Oramas, M., Ghodrati, A.,Tuytelaars, T: Modeling video evolution for action recognition(Conference Paper). In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol.7, pp. 5378–5387 (2015)
17. Qixiang, S., Ning, H., Congcong, Z., Shengjie, L.: Human skeleton action recognition method based on lightweight graph convolution. Comput. Eng.. Eng. **48**(5), 306–313 (2022)