Hindawi
Scientific Programming
Volume 2022, Article ID 2721618, 10 pages
https://doi.org/10.1155/2022/2721618

Research Article

Improved GCN Framework for Human Motion Recognition

Fen Zhou, Xuping Tu, Qingdong Wang, and Guosong Jiang

College of Computer Science, Huanggang Normal University, HuangGang, Hubei 438000, China

Correspondence should be addressed to Guosong Jiang; jiangguosong@hgnu.edu.cn

Received 25 March 2022; Revised 15 April 2022; Accepted 22 April 2022; Published 9 May 2022

Academic Editor: Jie Liu

Human recognition models based on spatial-temporal graph convolutional neural networks have been gradually developed, and we present an improved spatial-temporal graph convolutional neural network to solve the problems of the high number of parameters and low accuracy of this type of model. The method mainly draws on the inception structure. First, the tensor rotation is added to the graph convolution layer to realize the conversion between graph node dimension and channel dimension and enhance the model's ability to capture global information for small-scale tasks. Then the inception temporal convolution layer is added to build a multiscale temporal convolution filter to perceive temporal information under different time domains hierarchically from 4-time dimensions. It overcomes the shortcomings of temporal graph convolutional networks in the field of joint relevance of hidden layers and compensates for the information omission of small-scale graph tasks. It also limits the volume of parameters, decreases the arithmetic power, and speeds up the computation. In our experiments, we verify our model on the public dataset NTU RGB + *D*. Our method reduces the number of the model parameters by 50% and achieves an accuracy of 90% in the CS evaluation system and 94% in the CV evaluation system. The results show that our method not only has high recognition accuracy and good robustness in human behavior recognition applications but also has a small number of model parameters, which can effectively reduce the computational cost.

1. Introduction

Computer vision technology is a key link in the realization of artificial intelligence, and its emergence has given artificial intelligence great potential in visual perception. Among them, human action recognition is the most challenging technology in computer vision. The implementation of this technology can add to intelligent applications such as pedestrian following and behavior analysis. The most widely researched human behavior recognition methods are based on the human skeleton and, of course, image-based human behavior recognition methods. Human skeleton-based and image-based approaches are very different [1–4]. Skeleton-based methods take human skeleton recognition data as input, focus on analyzing the depth, spatial and temporal information of human skeletal joints, and then combine all features to achieve a behavior prediction result. Compared with image-based methods, human skeletal data are denser and reduce the computational cost by replacing a large number of pixel points with dense skeletal data. The skeleton-based action recognition method also performs better in the working environment of multiple targets and complex backgrounds [5–7].

Traditional skeleton-based methods rely on skeletal joint trajectories, with all method models based on recurrent neural networks of skeletal joint point data [8–11]. Some researchers would prefer to adopt deep neural network models, for which there is already relevant literature demonstrating their substantial advantages as well as their shortcomings. The skeletal data distribution is rather fragmented, and each skeletal joint point data are not locally linked. Therefore, for deep neural networks, a separate neural network needs to be tailored to accommodate the structured skeletal data to coordinate with all the skeletal joint point data [12, 13].

In the deep recurrent network model, only the connections in the feature point space can be analyzed, and the connections between features at the temporal level cannot be obtained. To solve this problem, researchers in the literature [14, 15] used a long short-term memory network (LSTM)

[16] to feature extraction from skeletal data, where the authors first divided the skeletal data into slices, each corresponding to an individual LSTM unit, and merged all. Such an architectural design improves the model's spatial perception of the skeletal data. However, this method suffers from the manual architecture predetermined to the rule limitation, which reduces the generalization ability and robustness of the network [17]. Considering the spatial and temporal features of skeletal data, the literature [18] introduced a graph convolutional network, which breaks the limitation of 2-dimensional data and can handle any graph structure. It transposes the computation of graph convolutional network to skeletal nodal data, which can dimensionally integrate the connections between spatial feature points. The literature [19] presented a spatial-temporal graph convolutional neural network based on the previous study, which can represent each skeletal data point in a graph structure and then perform feature extraction in a graph convolutional pattern as a way to obtain the spatial features between skeletal joint points [20–22]. In addition, the model adds a temporal convolution unit to integrate the temporal links between skeletal joint points, estimate the trajectory of skeletal joints, and finally predict the class of behavior [23].

Based on preliminary research and experiments, this paper proposes the Inception-ST-GCN (IST-GCN) method, which aims to reduce the complexity of building the neural network architectures while capturing the global information of the graph. In this paper, a tensor rotation module will be added to rotate the graph dimension to the RGB dimension and use the one-dimensional convolution Conv 1×1 to capture the global information afterward. A new inception layer multiscale temporal convolution filter is added to divide it into four branches with different temporal perception domains to capture richer temporal information and greatly decrease the volume of model parameters. The IST-GCN method achieves a compact and efficient network. To test the effectiveness of the method in this paper, we perform experimental validation on the public dataset NTU RGB + D. The results show that the number of parameters of the method in this paper is greatly reduced compared with the original ST-GCN model, and the accuracy and precision of action recognition are greatly improved.

The remainder of this paper is laid out as follows: Section 2 introduces the construction of the basic network and the principles of mathematical equations. Section 3 details the principles and implementation procedures related to the improved human action recognition network. Section 4 presents the relevant experimental datasets and analysis of the results. Finally, Section 5 reviews our findings and reveals some additional research.

2. Basic Network

Through our preliminary examination, we apply the graph convolutional neural network as the base network, and its network structure is shown in Figure 1. This network is an upgrade for the graph convolutional network, aiming to optimize the perceptual domain of the graph convolution and increase the joint of the graph convolutional network for the feature relations at the temporal level. The main purpose of this network is sequence encoding the skeletal data and predict the joint behavior by the spatial features and temporal associations between skeletal joint points. For skeletal feature acquisition, we usually use the OpenPose [24] algorithm to localize the human body using 25 skeletal points and the connections among different skeletal points as human joints. The input is usually a video sample in AV format, and each frame of the sample video corresponds to this set of joint coordinates. The OpenPose algorithm can split and resolve each set of joint coordinates and map them to each skeletal unit map node of the human body, using the joints and the edges of the human body as boundaries to build a complete spatial-temporal map. In other words, the input of OpenPose can also be understood as a set of joint coordinates of skeletal points in the same way as the 2-dimensional pixel intensity vector input of the convolutional neural network. To obtain a wider range of information, the graph convolutional network is then stacked and all outputs are then fed into the classifier in parallel.

The input in Figure 1 is a fixed skeleton sequence, assuming that T represents the constituent sequence of the total number of skeletons, V represents the number of skeletal joints, and $G = (N, E)$ denotes the set of constructed skeleton spatial-temporal sequences, where $N = \{v_{ti} | t = 1, ..., T, i = 1, ..., V\}$ traverses the skeleton joints obtained along with all-time sequences, and v_{ti} denotes all nodes. E denotes the set of connections between joints, and consists of E_T and E_S. An arbitrary human skeleton joint (i, j), $E_S = \{(v_{ti}, v_{tj}) | i, j = 1..., V, t = 1, ..., T\}$ denotes the composition of skeleton intra-joint connections within time t. The subset of intra-skeletal connections E_S is divided into K disjoint regions in the center of gravity rule and is represented using the adjacency matrix encoding $\widetilde{A}_k \in \{0, 1^{V \times V}\}$. $E_T = \{(v_{ti}, v_{(t+1)i}) | i = 1..., V, t = 1, ..., T\}$ denotes the union of connections between all skeletal joints in a continuous time series. The fusion of the above features results in a sequence diagram that can be extended in the spatially mapped temporal dimension.

The literature [25] optimized the spatial submodule of the spatial-temporal graph convolutional neural network and proposed the following graph convolution equation:

$$
\begin{aligned}
f_{\text{out}} &= \sum_{k}^{K_s} (f_{in} A_k) W_k, \\
A_k &= D_k^{(-1/2)} \left(\widetilde{A}_k + I\right) D_k^{(-1/2)}, \\
D_{ii} &= \sum_{k}^{K_s} \left(\widetilde{A}_k^{ij} + I_{ij}\right),
\end{aligned}
\tag{1}
$$

where $\widetilde{A}_s$ denotes the adjacency matrix of internal connections of skeletal nodes, I denotes the unit matrix, K_s denotes the size of the convolution kernel in spatial dimensions, and W_k denotes the training weights. The temporal convolution module is $1 \times K_t$. In 2D graph convolution, and the perceptual field of the convolution kernel is not considered when operating (C_{in}, V, T) in the (V, T) dimension, where K_t denotes the number of frames.

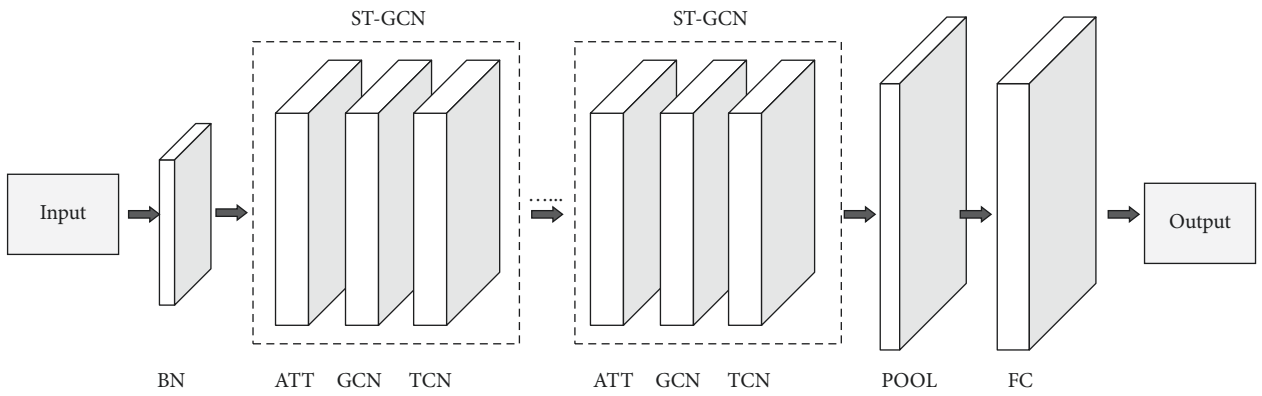

Figure 1: ST-GCN network.

The graph structures in graph convolution are predefined, and to increase their adaptability, the literature [26] uses a fixed adjacency matrix and proposes an adaptive graph convolution formula as follows:

$$f_{\text{out}} = \sum_{k}^{K_s} f_{in}(A_k + B_k + C_k)W_k, \tag{2}$$

where B_k denotes the parameters learned in training and C_k denotes the connected vertices determined with the over-similarity function.

3. Improved Action Recognition Network

The spatial-temporal graph convolution model uses a predefined structural graph as a topological constraint to achieve the ability of different time-step graphs to share the same topology, and such a structure leads to the inability of the graph task to fully capture the relevant features of the hidden joint layer. To solve this problem, our most common approach is to build a regional neural network using a local perceptual domain as the starting point and a small-scale graph task in the experimental region. This can easily produce global information omission. To simulate the principle of computation of pixel points by convolutional neural networks, each graph node and adjacent graph nodes become the key nodes for graph convolution computation in the graph convolution task. Considering the problem of density heterogeneity and narrow local structure between neighboring nodes, in our improved network, we employ node features of fixed size for feature learning in the temporal dimension, selectively ignoring the size of cluster features, and being able to capture more features in the temporal dimension. Therefore, we present the inception spatial-temporal graph convolutional network (IST-GCN), which applies the inception structure to some network layers as a way to reduce the model parameters, broaden the network width, and enhance the robustness of the model.

3.1. Inception Module. The inception module is a sparsity structure proposed in 2015, which has excellent feature expression capability and local topology capability. When the image is input, the pixel point population is involved in a series of convolution operations and pooling operations to obtain features at different scales from different scales of convolution kernels. All the output results are taken for parallel processing to filter out the best image features. The original structure of inception is shown in Figure 2. Its network structure mainly contains three scales of convolutional kernels and a 3×3 pooling, through which a combination of 1, 3, and 5 convolutional kernels can fully acquire large-scale sparse features and small-scale nonsparse features. Such structures not only increase the network width but also increase the adaptability of the network to different scales. Finally, all features are synthesized by a concatenation operation to obtain the nonlinear properties of the features.

3.2. Graph Convolutional Layer Improvement Strategy. Our proposed IST-GCN model originates from a two-part optimization of the spatial-temporal graph convolutional network. The first part is to optimize the graph convolutional network layers; the second part is to add the inception layer. In the graph convolutional layer, the original model aims to obtain spatial location information between the human skeletal joint points to achieve the representation of the joint points. It should start from the initial neighboring nodes to build up a local perceptual domain, in which a large number of sample nodes are generated. Although many false samples are generated at this point, adding topological angle restrictions in the subsequent process of filtering the sequence in Euclidean space can filter the false samples. When all sample nodes are in Euclidean space, all sample nodes can be considered as point from the global level view, and the sequence of points is considered as a one-dimensional vector. In this case, to capture a large number of sample node features, a large-scale graph convolution sum is required, whose size is consistent with the number of nodes. To properly solve this problem, we propose a tensor rotation strategy. We add a tensor rotation module, which we call Rotate tensor GCN (R-GCN), at the beginning and end of the graph convolution layer. The detailed network structure is shown in Figure 3.

By the action of the tensor rotation module, each sample node can share the same set of identical topological matrices,

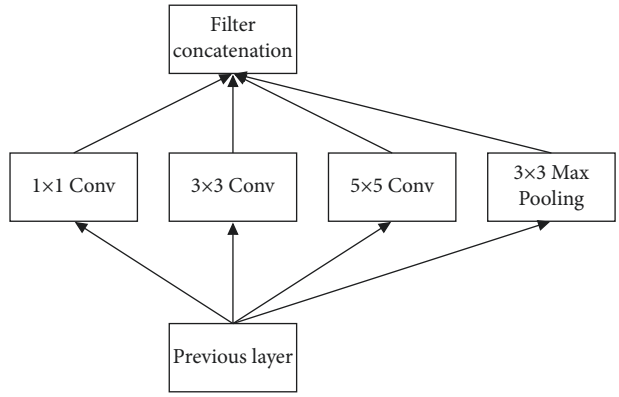

Figure 2: Inception network.

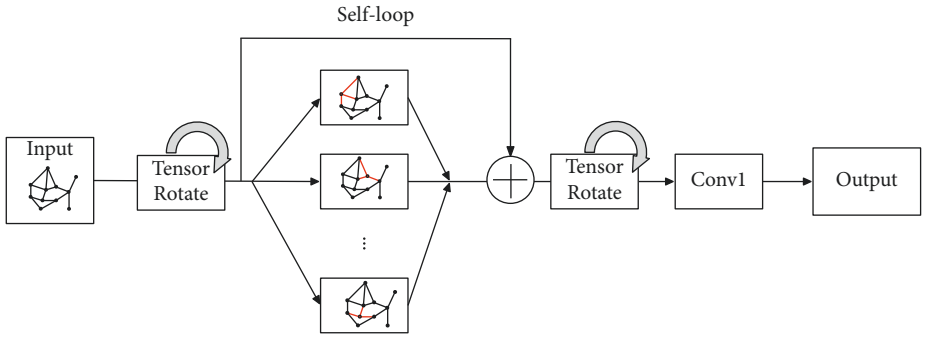

Figure 3: R-GCN network.

and all nodes can participate in the process of capturing global information. Taking human nodes as an example, each graph contains 25 nodes, and in the fully connected layer, we choose a filter of size 25. The rotation tensor module will rotate a tensor according to the different nodes separately so that the dimensionality of the nodes and the dimensionality of the channels remain the same. By tensor rotation, the predefined topological matrix is discarded and the global features are learned adaptively according to the self-cycling unit for joint relevance. Finally, the global information is integrated through the tail-Conv 1×1 dimensionality reduction. Such a structural design can effectively reduce the use of higher-order polynomial estimation layer by layer to capture higher-order features, thus achieving a reduction in the number of parameters.

3.3. Inception Layer Design Strategy. We consider using the inception structure to broaden the spatial-temporal graph convolutional network because of the sparse structure advantage of inception. More feature information can be obtained by the layout of the sparse structure while avoiding the increase in the number of parameters. We refer to the optimization process of inception from V1 to V4 and discover the one-dimensional convolutional dimensionality reduction method [27–29]. We are building the inception time convolution network (I-TCN), and the expansion of parameters is exacerbated by the exponentially growing expansion coefficients in the time convolution layer to widen the network. In contrast, the inception tiling structure is incremented according to layers, and each branch is preceded by adding Conv 1×1 dimensionality reduction to assign different expansion settings to each branch, allowing the time-scale information to be graded into the inception branches and achieving information integration in different time dimensions. Through the above structure of time coefficient assignment, the exponential growth of coefficients is avoided and the purpose of reducing the number of parameters is achieved.

Two two-layer I-TCN layers are added at the end of each IST-GCN cell, and the TCN is divided into 4 branches according to the hierarchical principle, with each branch producing output to the corresponding group, whose structure is shown in Figure 4. The initial value of the expansion coefficient n of the network is 1. As the network deepens, the layer units increase step by step, and the maximum value of the expansion coefficient is 4. This external connection refers to the residual structure, which passes through a one-dimensional convolution with a step size of 2 in the middle, and this design can avoid the gradient dispersion problem. Improving the temporal convolutional network by inserting the inception structure allows for capturing more time-scale information while reducing the number of network parameters by a large amount and reducing the computational cost. A compact and efficient temporal feature extraction network is achieved by using different temporal filters to adaptively select the best feature information to optimize the classification problem.

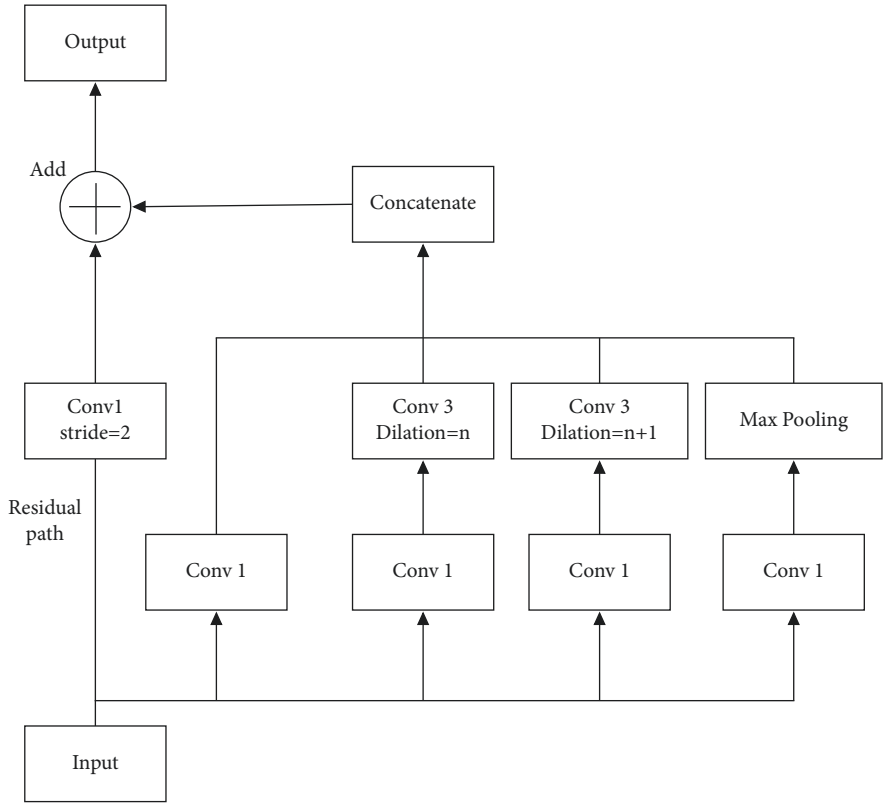

Figure 4: I-TCN network.

3.4. IST-GCN Action Recognition Process. The process of human action recognition based on the IST-GCN model is shown in Figure 5. Firstly, the sample video data are input, and the video data are processed in frames during the analysis process. The human joints under different frames have the problem of position change, but the set of all joint points in different frames obeys random distribution. Therefore, we first select the batching standard module (BN) in the first layer of the network hierarchical distribution to normalize the joint point data at the temporal level and spatial level to make the input skeletal data more standardized, reduce the error volatility, and optimize the algorithm's convergence. In the second layer of the network, we choose the attention mechanism (ATT), which connects our new R-GCN layer and the I-TCN layer in the next network. The R-GCN layer relies on the tensor rotation operation to obtain global information, after which the obtained global features are input into the I-TCN to analyze the linkage relationship among the nodal features at the temporal level, supplemented by the ATT mechanism to weaken the non-conforming features that do not conform to the bounded range of the model and filter features of different time-scales. The whole network consists of nine IST-GCN units sequentially connected to fully capture and fuse the graph feature information, then perform average pooling, then classify the features through the fully connected layer, and finally output the behavior prediction results according to the classification weights.

4. Experiment

4.1. Datasets. To validate the performance of our method, we chose the public dataset NTU RGB + *D* [30] for experimental test validation. This dataset is one of the more comprehensive datasets covering categories in human action recognition studies. The dataset contains a total of three types of production specifications, which are the two-person interaction dataset, the medical interaction dataset, and the daily interaction dataset. It can be subdivided into 60 categories of actions based on action types, with a total of 56880 sample sequences. All videos are stored in a uniform dataset standard, and the maximum video frames of each sample video do not exceed 300 frames. At the same time, all sample data are preprocessed by OpenPose human skeleton detection, and the corresponding skeleton data and Jason files are stored separately. In addition, a set of independent evaluation criteria, namely, Cross-Subject (CS) and Cross-View (CV), is proposed for this dataset. The CS evaluation system is evaluated based on the ID number of the person in the dataset as a sequence, and the CV evaluation system is evaluated based on the camera ID number as a sequence. The detailed volume of the training and testing datasets are shown in Table 1.

4.2. Experimental Details. In the action recognition experiments, we mainly focus on action jogging as the control standard to verify whether the action recognition results

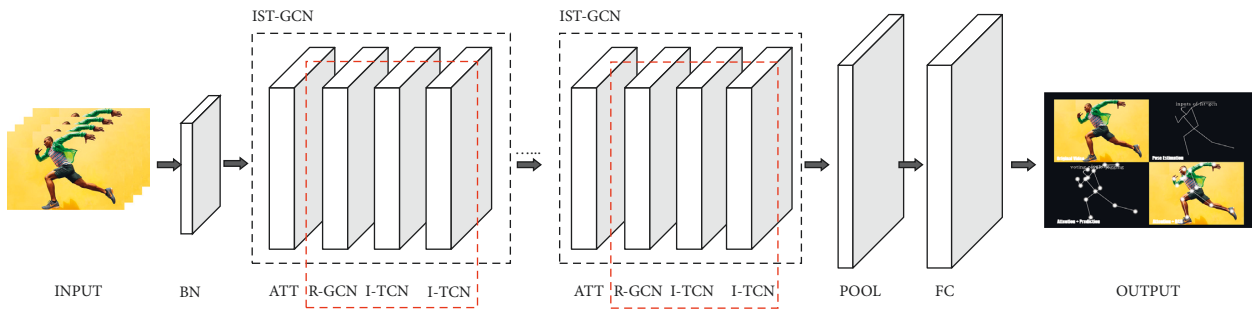

Figure 5: IST-GCN human action recognition process.

Table 1: The detailed information of datasets.

	NTU RGB + *D*	
	CS	CV
Train	40320	37920
Test	16560	18960
Total	56880	56880

match with the real action, each test sample is 300 frames, while the experiments are divided into single-player action recognition experiments and multiplayer action recognition experiments to test the performance of the improved method hierarchically while comparing with the spatial-temporal map convolutional neural network model.

4.2.1. Single Action Recognition Experiment. The performance of single-person recognition result is shown in Figure 6, it can be seen that the action recognition result matches with the experimental preset result, the effect is better and the action recognition result is accurate.

Compared with the spatial-temporal map convolutional neural network model, the single-person action recognition effect is not much different, and the comparison experiment is shown in Figure 7. Although there are a few frames that recognize the action as a triple jump and occasionally misrecognition occurs, the final score voting result still matches the real action and has little impact on the overall action recognition result.

4.2.2. Multiplayer Action Recognition Experiment. The performance of multiperson recognition result is shown in Figure 8, which shows that the action recognition effect is good, and a few frames appear to misrecognition situation, but it does not affect the overall action recognition, and the recognition result is accurate.

Compared with the original spatial-temporal map convolutional neural network model, the recognition effect of our method is superior, and the comparison of the action recognition effect is shown in Figure 9.

As shown in Figure 9 Experiment A, two-thirds of the frames of the original ST-GCN method identify the action as triple jump, although there are also some frames identified as real action jogging, but the overall triple jump action score is higher, so the final action recognition result is triple jump. Our method uses different scales of time windows to capture information and has better control of global information, so it performs well in the multiperson action recognition experiment and the recognition results are accurate. From Figure 9, experiment B in the recognition effect of the original ST-GCN algorithm, one person was obscured and although the skeletal information was recognized, the action could not be classified, and then the overall action was recognized as roller skating, which could not be matched with the real action. The effect of multiperson action recognition experiments is not as good as that of single-person recognition experiments. The more the number of people, the lower the accuracy of human skeleton recognition and the efficiency of action classification. We try to control the multiperson action recognition experiment to less than three people in the experiment. Our method can recognize and correctly categorize the occluded part of the action, further highlighting the advantages of our proposed IST-GCN method.

4.3. Experimental Results Analysis. Our proposed IST-GCN method involves the improvement of two main parts, namely, the rotated tensor module in the graph convolution layer (R-GCN) and the inception structure embedding in the temporal convolution layer (I-TCN). To verify the effect of each, ablation experiments were performed. First, the GCN in ST-GCN was replaced with R-GCN, and the group of experiments was named with the letter *R* to construct the R-GCN efficiency testing experimental group. Secondly, the TCN in ST-GCN was replaced with I-TCN, and the group was named with the letter I. The experimental group was constructed to verify the performance of the I-TCN module. The above two groups were validated with the spatial-temporal map convolutional neural network and our proposed IST-GCN on the NTU RGB + *D* dataset. The results were compared in terms of accuracy (Acc), bone recognition accuracy (Bone), joint recognition accuracy (Joint), and number of parameters (Param) levels as shown in Table 2.

The R-GCN technique improves overall accuracy by 3.7 percent, and the number of parameters is lowered proportionally, as shown in Table 2. The I-TCN approach improves overall accuracy by 7.5 percent and reduces the number of parameters by half. The results reveal that I-TCN has a greater impact on overall performance than R-GCN, although less effective than I-TCN in terms of overall

Figure 6: Single-person action recognition.

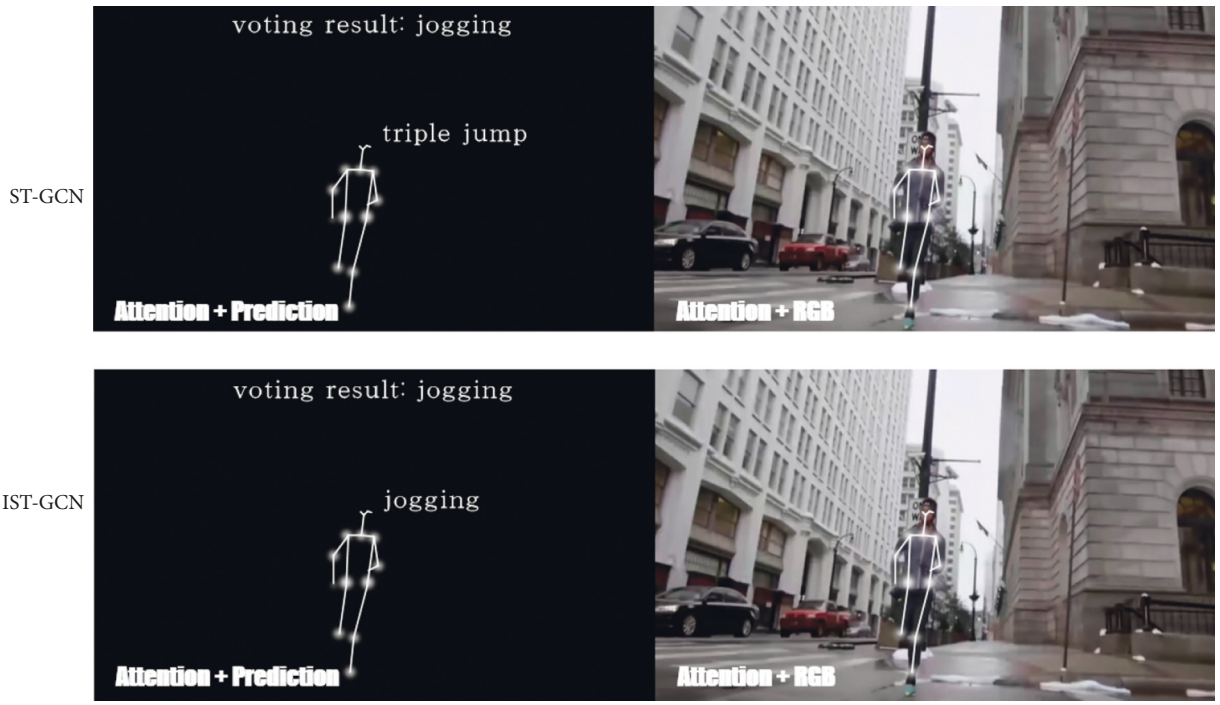

Figure 7: Comparison of ST-GCN and IST-GCN single-person action recognition effects.

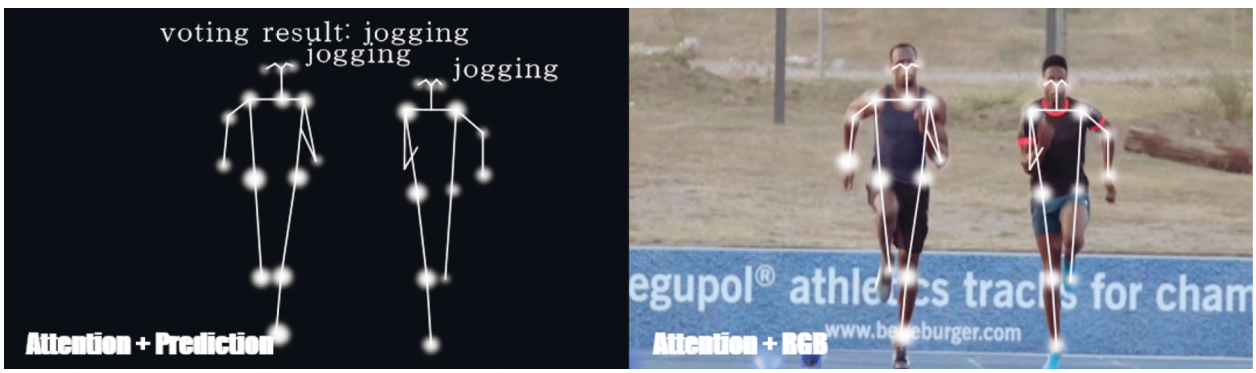

Figure 8: Multiperson action recognition.

performance improvement, is indispensable in capturing the global feature level. The two optimizations mirror each other and prove the effectiveness of the IST-GCN method.

To verify the effectiveness of our IST-GCN, we compare four different kinds of skeleton-based action recognition models, dynamic skeleton [31], ST-GCN [18], P-LSTM [30]

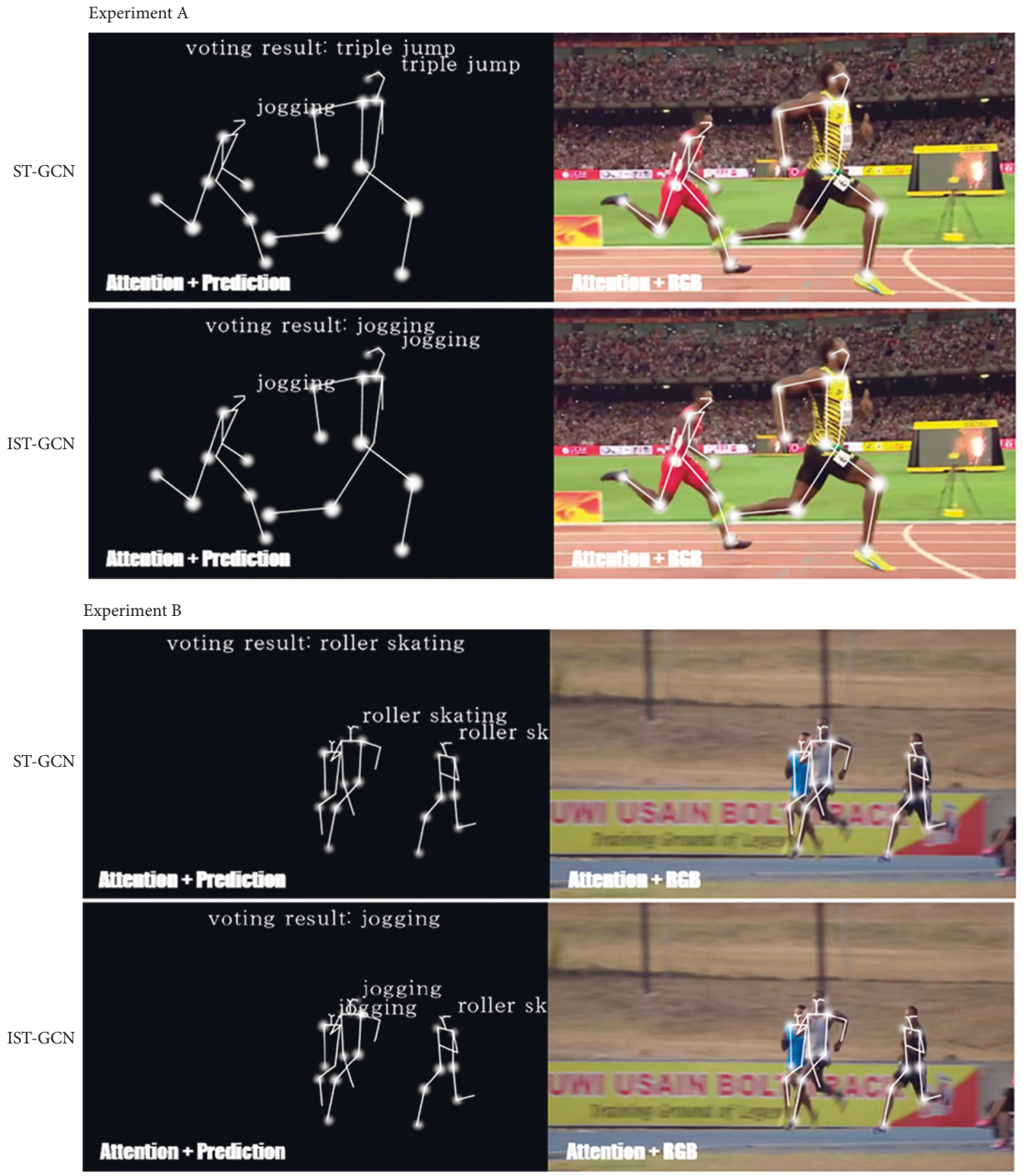

Figure 9: Comparison of multiperson action recognition effects between ST-GCN and IST-GCN.

Table 2: Results of ablation experiments.

Method	Joint(%)	Bone(%)	Acc(%)	Param(M)
ST-GCN	79.1	79.8	80.1	3.14
R	82.9	83.2	83.8	2.36
I	86.8	87.1	87.6	1.61
Ours	88.9	89.5	90.2	1.36

Table 3: Comparison of our method and different types of action recognition methods.

Method	CS(%)	CV(%)
Dynamic skeleton	60	64
P-LSTM	63	71
TCN	74	83
ST-GCN	81	88
Ours	90	94

Table 4: Comparison of our method and similar optimized action recognition methods.

Method	Params(M)	Acc (%)
ST-GCN	3.03	81
AS-GCN	4.21	84
2S-AGCN	3.52	87
NAS-GCN	6.62	87
Shift-GCN	7.34	95
Ours	1.43	91

and TCN [32]. The dynamic skeleton represents a series of action recognition models based on hand-crafted labels, P-LSTM denotes a series of recurrent neural network classes, TCN denotes a series of convolutional neural network classes, and ST-GCN denotes a series of hands-on models based on graph convolutional neural networks. The above four methods and our method are validated on the NTU RGB + *D* dataset. The experimental data is shown in Table 3.

The experimental comparison results in Table 3 indicate that in the validation experiments of the dataset NTU RGB + *D*, the GCN-based action recognition method greatly outperforms other types of action recognition methods, proving that graph convolutional networks have great advantages. Our method compared with the spatial-temporalspatial-temporal graph convolutional neural network model improves the accuracy in CS metrics by 9%, reaching 90% and in CV metrics by 6% and reaching 94%.

To verify the effectiveness of our method among similar optimization methods for graph convolutional neural networks, we compared four algorithms that perform better among current variant methods for graph convolutional neural networks in terms of both number of parameters (Params) and accuracy (Acc), namely AS-GCN [33], 2S-AGCN [26], NAS-GCN [34], and Shift -GCN [35]. The validation was carried out in dataset NTU RGB + *D* with CS evaluation metrics, and the comparison results are shown in Table 4.

Table 4 reveals the findings of the experimental comparison. The comparison results between AS-GCN, 2S-AGCN, and NAS-GCN under the evaluation index of CS indicate that our method has better efficiency with an accuracy of 91%, both in the number of model parameters and accuracy. Given the Shift-GCN method, which introduces a more complex hyperbolic space structure, the classification accuracy is further optimized. Even though the accuracy is not as good as that of Shift-GCN, the number of model parameters in this paper adopts the inception structure to form a more compact model, and the number of model parameters in our improved method is only one-fifth of that of the Shift-GCN method, which greatly decreases the computational cost. Furthermore, there are fewer parameters in this model than in previous ones. All of this demonstrates the efficacy of our strategy.

5. Conclusion

In this paper, we present a deep learning method for human action recognition based on the IST-GCN framework, which optimizes the recognition accuracy of the model by reducing the model parameters. First, we add a tensor rotation module in the graph convolution layer to better capture the global features of the graph task. Then we add the inception structure in the temporal convolution layer to build a multiscale temporal convolution filter to obtain temporal information in different temporal perceptual domains and reduce the arithmetic power. Finally, we perform experimental validation on the public dataset NTU RGB + *D*. The accuracy of CS evaluation reaches 90% and the accuracy of CV evaluation reaches 94%. The results reveal that our optimized method is robust and accurate, which not only improves the efficiency of the graph topology learning process but also greatly decreases the volume of parameters. Compared with the spatial-temporalspatial-temporal graph convolutional neural network model and similar graph convolutional optimization algorithms, the advantages of our method are outstanding.

As can be seen from the experimental results in Table 4, there is still a certain gap between the accuracy of our method and the Shift-GCN. Although we have a clear advantage in the number of parameters, accuracy is always the first assessment index as the effect of human action recognition. In the next work, we will consider using hyperbolic spatial structure to optimize the accuracy, and also ensure that the volume of parameters is small. To achieve a human action recognition model with high accuracy, few parameters, high robustness, and good stability.

Data Availability

The dataset can be accessed upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] C. Chen, R. Jafari, and N. Kehtarnavaz, "Utd-Mhad: a multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP)*, pp. 168–172, IEEE, Canada, 2015.

[2] Z. Jia, Y. Lin, J. Wang et al., "Multi-view spatial-temporal graph convolutional networks with domain generalization for sleep stage classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 1977–1986, 2021.

[3] A. Assadzadeh, M. Arashpour, I. Brilakis, T. Ngo, and E. Konstantinou, "Vision-based excavator pose estimation

using synthetically generated datasets with domain randomization," *Automation in Construction*, vol. 134, p. 104089, 2022.

[4] H. Wang, Z. Xie, L. Lu, L. Li, and X. Xu, "A computer-vision method to estimate joint angles and L5/S1 moments during lifting tasks through a single camera," *Journal of Biomechanics*, vol. 129, p. 110860, 2021.

[5] P. M. Griffin, "3-D object pose determination using computer vision," *Computers & Industrial Engineering*, vol. 19, no. 1-4, pp. 215–218, 1990.

[6] Z. Jia, Ji Junyu, and X. Zhou, *Hybrid Spiking Neural Network for Sleep EEG Encoding[J]*, Science China Information Sciences, vol. 65, no. 4, pp. 1–10, 2022.

[7] G. Morinan, Y. Peng, S. Rupprechter et al., "Computer-vision based method for quantifying rising from chair in Parkinson's disease patients," *Intelligence-Based Medicine*, vol. 6, p. 100046, 2022.

[8] Z. Jia, X. Cai, and Z. Jiao, "Multi-modal physiological signals based squeeze-and-excitation network with domain adversarial learning for sleep staging," *IEEE Sensors Journal*, vol. 22, no. 4, pp. 3464–3471, 2022.

[9] Z. Jia, Y. Lin, Y. Liu, Z. Jiao, and J. Wang, "Refined non-uniform embedding for coupling detection in multivariate time series," *Physical Review A*, vol. 101, no. 6, Article ID 062113, 2020.

[10] X. Liu, Z. You, Y. He, S. Bi, and J. Wang, "Symmetry-Driven hyper feature GCN for skeleton-based gait recognition," *Pattern Recognition*, vol. 125, Article ID 108520, 2022.

[11] T. Huynh-The, C.-H. Hua, N. Anh Tu et al., "Hierarchical topic modeling with pose-transition feature for action recognition using 3D skeleton data," *Information Sciences*, vol. 444, pp. 20–35, 2018.

[12] D. Das Dawn and S. H. Shaikh, "A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector," *The Visual Computer*, vol. 32, no. 3, pp. 289–306, 2016.

[13] Z. Jia, X. Cai, Y. Hu et al., "Delay propagation network in air transport systems based on refined nonlinear Granger causality," *Transportmetrica B: Transport Dynamics*, pp. 586–598, 2022.

[14] X. Jiang, F. Zhong, Q. Peng, and X. Qin, "Online robust action recognition based on a hierarchical model," *The Visual Computer*, vol. 30, no. 9, pp. 1021–1033, 2014.

[15] J. Wu, D. Hu, and F. Chen, "Action recognition by hidden temporal models," *The Visual Computer*, vol. 30, no. 12, pp. 1395–1404, 2014.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] X. Shen and Y. Ding, "Human skeleton representation for 3D action recognition based on complex network coding and LSTM," *Journal of Visual Communication and Image Representation*, vol. 82, Article ID 103386, 2022.

[18] L. Chaolong, C. Zhen, and Z. Wenming, "Spatio-temporal Graph Convolution for Skeleton Based Action recognition," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, LO, USA, 2018.

[19] S. Yan, Y. Xiong, and D. Lin, "Spatial-temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the Thirty-second AAAI conference on artificial intelligence*, New Orleans, LO, USA, 2018.

[20] T. Alsarhan, U. Ali, and H. Lu, "Enhanced discriminative graph convolutional network with adaptive temporal modelling for skeleton-based action recognition," *Computer Vision and Image Understanding*, vol. 216, p. 103348, 2022.

[21] Y. Li, D. Ma, Y. Yu, G. Wei, and Y. Zhou, "Compact joints encoding for skeleton-based dynamic hand gesture recognition," *Computers & Graphics*, vol. 97, pp. 191–199, 2021.

[22] Y. Liu, H. Zhang, D. Xu, and K. He, "Graph transformer network with temporal kernel attention for skeleton-based action recognition," *Knowledge-Based Systems*, vol. 240, p. 108146, 2022.

[23] C. Ding, S. Wen, W. Ding, K. Liu, and E. Belyaev, "Temporal segment graph convolutional networks for skeleton-based action recognition," *Engineering Applications of Artificial Intelligence*, vol. 110, p. 104675, 2022.

[24] Z. Cao, T. Simon, and S. E. Wei, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7291–7299, Honolulu, HI, USA, 2017.

[25] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," arXiv preprint arXiv: 1609.02907, 2016.

[26] L. Shi, Y. Zhang, and J. Cheng, "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, Long Beach, CA, USA, 2019.

[27] S. M. Sam, K. Kamardin, and N. N. A. Sjarif, "Offline signature verification using deep learning convolutional neural network (CNN) architectures GoogLeNet inception-v1 and inception-v3[J]," *Procedia Computer Science*, vol. 161, pp. 475–483, 2019.

[28] F. Chen, J. Wei, B. Xue, and M. Zhang, "Feature fusion and kernel selective in Inception-v4 network," *Applied Soft Computing*, vol. 119, p. 108582, 2022.

[29] W. Abdul, M. Alsulaiman, S. U. Amin et al., "Intelligent real-time Arabic sign language classification using attention-based inception and BiLSTM," *Computers & Electrical Engineering*, vol. 95, p. 107395, 2021.

[30] A. Shahroudy, J. Liu, T. T. Ng, and G. Wang, "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, Las Vegas, NV, USA, 2016.

[31] J. F. Hu, W. S. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for RGB-D activity recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5344–5352, Massachusetts, MA, USA, 2015.

[32] T. S. Kim and A. Reiter, "Interpretable 3d human action analysis with temporal convolutional networks," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1623–1631, IEEE, Honolulu, HI, USA, 2017.

[33] M. Li, S. Chen, and X. Chen, "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, Long Beach, CA, USA, 2019.

[34] W. Peng, X. Hong, H. Chen, and G. Zhao, "Learning graph convolutional network for skeleton-based human action recognition by neural searching," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2669–2676, New York, NY, USA, 2020.

[35] K. Cheng, Y. Zhang, X. He, C. Weihan, C. Jian, and L. Hanqing, "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, Nashville, TN, USA, 2020.