Algorithm for Human Abnormal Behavior Recognition Based on Improved Spatial Temporal Graph Convolutional Networks

Qi Wu[1], Xiaoyan Zhao[1,2], Zhaohui Zhang[1,2(✉)], Tianyao Zhang[1,2], and Zexuan Peng[1]

[1] School of Automation and Electrical Engineering, University of Science and Technology, Beijing, 30# Xueyuan Road, Haidian District, Beijing 100083, China
{Zhaoxiaoyan,zhangzhaohui}@ustb.edu.cn

[2] Beijing Engineering Research Center of Industrial Spectrum Imaging, University of Science and Technology, Beijing, 30# Xueyuan Road, Haidian District, Beijing 100083, China

Abstract. With the increasing demand for public safety, the field of abnormal human behavior recognition has undergone significant development. In addressing the low accuracy issue of existing abnormal behavior recognition algorithms due to factors such as environmental influences, changes in viewpoint, and scale variations, this study proposed an improved Spatial temporal graph convolutional network. By incorporating spatial attention and channel attention mechanisms at relevant positions in the network, a dynamic optimization of the skeletal structure graph of the human body was achieved. This ensured that key nodes expressing motion information in the skeletal graph received greater weight values, ultimately improving the accuracy of abnormal behavior classification. To this end, an abnormal behavior dataset was constructed and transformed into skeletal information recognizable by the proposed algorithm using OpenPose. Extensive experiments were conducted on this dataset as well as the large-scale NTU RGB + D dataset using the improved algorithm. The results demonstrate that the algorithm has achieved an increase of approximately 5% in recognition accuracy compared to its pre-improvement state, placing it among the top-performing algorithms in various comparative evaluations.

Keywords: Anomalous Behavior Recognition · Graph Convolutional Network · Attention Mechanism

1 Introduction

As people's demand for and concern about public safety increased, computer vision for the field of security became a research hotspot [1]. Among them, human abnormal behavior recognition, as an important branch of security, has been widely applied in scenarios such as smart campuses, community security, and smart elderly care. Combining artificial intelligence with traditional video surveillance systems can quickly and accurately detect events that compromise public safety, thus better ensuring people's security and saving manpower and expenses, which are powerful guarantees for building a civilized city.

B. Xin et al. (Eds.): IWACIII 2023, CCIS 1932, pp. 29–42, 2024.
https://doi.org/10.1007/978-981-99-7593-8_4

Traditional skeleton-based human behavior recognition methods mainly rely on manually designed features [2]. These methods can achieve good recognition results in certain specific actions or scenes, but their generalization ability is poor [3], and feature extraction is complex. In recent years, with the continuous development of deep learning technology, deep learning models based on skeleton data, such as convolutional neural networks (CNN) [4], recurrent neural networks (RNN) [5], and graph convolutional neural networks (GCN) [6, 7], have been developed and applied. The method of using CNN models to process skeleton data involves transforming the skeleton data into pseudo-images and then inputting them into the network for recognition. RNN is more suitable for processing data with temporal sequences, while human skeleton data not only contains temporal information of the same node but also includes joint connection information of different nodes. Therefore, it often needs to be combined with CNN models to extract spatial features. The GCN-based method takes graph data as input, which means it contains topological graph structure data with key points and connection information. Human skeleton data meets such data requirements, so using GCN for skeleton-based human behavior recognition has inherent advantages.

In recent years, many researchers at home and abroad have applied GCN to the field of human behavior recognition: Shi et al. used first-order and second-order information of joints and bones to propose a skeleton-based dual-stream adaptive graph convolutional network (2s-AGCN) [8]. In reference [9], a graph attention network was designed, which employed stacked hidden self-attention layers to assign different weights to different nodes. Its performance was significantly superior to the contemporaneous RNN algorithms. YAN et al. pioneered the introduction of spatiotemporal characteristics into the GCN network and proposed the spatiotemporal graph convolutional network (ST-GCN) [10], which captures both time and space dimensions of information through spatiotemporal graph convolution to better recognize human actions. The above research fully utilizes the characteristics of GCN and combines the skeleton information of key points and joint connections to learn the features of human behavior in the temporal and spatial dimensions, and all have achieved good results. However, there are still some shortcomings. For example, the focus on different parts of the body is different in specific scenarios, which requires different weights for different nodes and their connection relationships in each frame of skeleton data. In addition, not all channels in the input feature channels composed of skeleton data are useful for recognizing targets. Dynamically ignoring unimportant channel information and focusing on useful channel features will inevitably improve the recognition effect.

Based on the above analysis, the main contribution of this study is as follows:

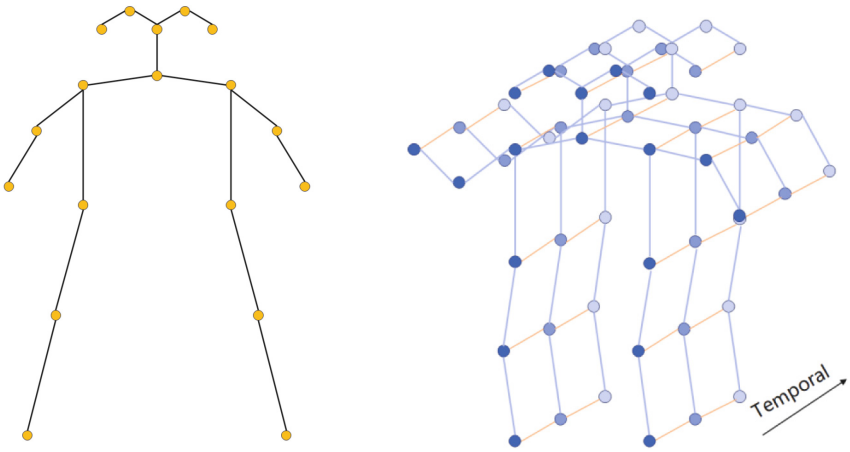
1) Designing an improved method to address the shortcomings of ST-GCN in feature extraction. This method introduces a channel attention mechanism to handle the case where all feature weights in the channel are equal. Additionally, a spatial attention module is incorporated to give more attention to the more important parts of the skeleton data.
2) Defining three types of abnormal behaviors, namely falling, fighting, and smoking, which have a significant impact on people's production and daily life. A corresponding dataset was created, and the dataset was processed into skeleton data using OpenPose. The improved ST-GCN algorithm mentioned above was then applied to the abnormal

behavior recognition task. Through verification and comparison, it was found that this algorithm achieved good results.

2 ST-GCN Model with Hybrid Attention Mechanism

2.1 Constructing Spatiotemporal Graph of Human Skeleton

After processing the human body through OpenPose, a sequence of key points is obtained, which contains the coordinate information of the keypoints and the natural connection information between them. Utilizing this information, we can construct the human joint spatiotemporal graph that serves as input for the ST-GCN model. The specific construction method is as follows: For a single frame of human skeleton data, the natural connections between the keypoints are used to form the spatial graph. In Fig. 1 (a) below, the 18 skeleton points extracted by OpenPose are represented by solid yellow dots. The lines connecting the keypoints represent the skeletal connections of the human body, which abstract the actions performed by the person. For the same keypoint in different frames, the corresponding keypoints from adjacent frames are connected in sequence. This process allows us to construct a human skeleton connection graph, with all the keypoints as the node set, and the connections between the nodes as the edge set. This is illustrated in Fig. 1 (b) below:



(a). Single-frame human skeleton graph. (b). Skeleton spatiotemporal graph.

Fig. 1. Schematic diagram of human skeleton.

Assuming Fig. 1(b) represents a skeleton spatiotemporal graph with N nodes and T frames, it can be mathematically represented by the following formula: $G = (V, E)$, The set of its nodes is denoted as $V = \{v_{ti}\, \hat{u}t = 1, \ldots, T, i = 1, \ldots, N\}$. v_{ti} represents the t th keypoint on the i th frame. v_{ti} contains the coordinate information and confidence score of the human body keypoints in the image. E is the set of edges in the skeleton spatiotemporal graph, consisting of two parts: E_S and E_t. $E_s = \{v_{ti}v_{tj}(i, j) \in H\}$ is the

collection of connecting edges between adjacent keypoints within a frame, as shown by the blue lines in Fig. 2. In the equation, H represents a set of naturally connected human body joints; $E_f = \{v_{ti}v_{(t+1)i}\}$ is the collection of connecting edges between the same keypoints in adjacent frames, as shown by the orange lines in Fig. 2. At this point, the spatiotemporal graph of the skeleton is constructed, which can be understood as a three-dimensional skeleton. It contains the skeletal information of human body movements over a period of time and serves as the input for subsequent ST-GCN processing.

2.2 Spatial Temporal Graph Convolutional Networks

The advantage of graph convolutional networks lies in their ability to handle non-Euclidean distance graph data, such as human skeletons, transportation networks, social networks, etc. ST-GCN enriches graph convolutional networks in both the temporal and spatial dimensions, allowing them to capture the spatiotemporal characteristics of human body movements from skeleton sequences, thereby enabling more accurate recognition of human actions [11].

Taking the spatial two-dimensional convolution of a typical Convolutional Neural Network (CNN) as an example, the convolution output for a specific position can be expressed in the following form:

$$f_{out}(x) = \sum_{h=1}^{K}\sum_{w=1}^{K} f_{in}(P(x,h,w)) \cdot w(h,w) \tag{1}$$

In this equation, the feature map f_{in} has a number of input channels denoted as c, The convolution kernel has a size $K \times K$, The function P represents the sampling function. The sampling is conducted within the neighborhood (h, w) of the region x. The weight function matrix is represented by w.

In an image, the sampling function $P(h, w)$ refers to the collection of neighboring pixels centered around the x pixel. In a topological graph, the set of neighboring pixels is defined as follows:

$$B(v_{ti}) = \left\{v_{tj} \,|d\left(v_{tj}, v_{ti}\right) \leq D\right\} \tag{2}$$

$d\left(v_{tj}, v_{ti}\right)$ refers to the shortest distance from v_{tj} to v_{ti}, and D is the threshold for the sampling distance. In this context, the threshold D for sampling is set to 1. Therefore, the sampling function is defined as:

$$P\left(v_{tj}, v_{ti}\right) = v_{tj} \tag{3}$$

In 2D convolution, the pixels within a neighborhood are arranged around the central pixel according to a certain rule, allowing convolution operations to be performed using predefined convolution kernels. Similarly, in graph data, the neighboring pixels obtained from the sampling function can be divided into different subsets based on different partitioning strategies. Each subset is then assigned a label. By applying mapping operation $l_{ti} : B(v_{ti}) \rightarrow \{0, \ldots, K-1\}$ to the adjacent regions of node v_{ti} and assigning different weight parameters to these adjacent regions, we can obtain a weight function w:

$$W\left(v_{ti}, v_{tj}\right) = W'\left(l_{ti}\left(v_{tj}\right)\right) \tag{4}$$

Applying the newly defined sampling function and weight function to Eq. (1), we obtain the redefined spatial graph convolution formula:

$$f_{out}(v_{ti}) = \sum_{vtj \in B(v_{ti})} \frac{1}{Z_{ti}(v_{tj})} f_{in}\left(P(v_{ti}, v_{tj})\right) \cdot w(v_{ti}, v_{tj}) \tag{5}$$

The normalization term $Z_{ti}(v_{ti}) = |\{v_{tk}\}| l_{ti}(v_{tk}) = l_{ti}(v_{tj})|$, which is equivalent to the basis of the corresponding subset, balances the contributions of different sub-regions to the output. By substituting Eqs. (3) and (4) into Eq. (5), we obtain:

$$f_{out}(v_{ti}) = \sum_{vtj \in B(v_{ti})} \frac{1}{Z_{ti}(v_{tj})} f_{in}\left(v_{t_{ij}}\right) \cdot w'(l_{ti}(v_{ti})) \tag{6}$$

In the temporal dimension, as each node is fixed to have two adjacent nodes, after performing the spatial graph convolution mentioned above, extracting temporal features can be achieved by applying a two-dimensional convolution to the output feature map in the temporal dimension. This eventually enables spatiotemporal graph convolution operations.

The above equations explain the working principle of the ST-GCN network. Regarding the partitioning strategy mentioned earlier, a spatial configuration partitioning method is employed, as illustrated in Fig. 2 below.

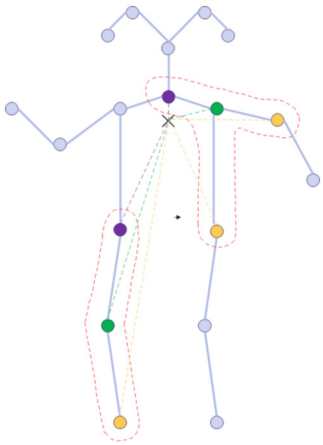


Fig. 2. The partitioning strategy for the human body skeleton diagram.

In the diagram, " ×" represents the centroid of the stock price. The red area denotes the neighborhood of green key nodes. In this partitioning strategy, subsets are formed based on the proximity of each node to the central point, resulting in three categories. The purple key nodes closest to the centroid are referred to as centripetal nodes, the green nodes represent the target nodes themselves, and the yellow nodes farthest from the centroid are referred to as centrifugal nodes. The division into these three subsets

reflects the concentric, centrifugal, and stationary motion characteristics of the human body. This can be expressed with the following equation:

$$l_{ti}(v_{ti}) \begin{cases} 0 & r_j = r_i \\ 1 & r_j < r_i \\ 2 & r_j > r_i \end{cases} \tag{7}$$

In this context, r_i denotes the distance between the target node and the centroid of the skeletal structure. Similarly, r_j represents the distances of each key node from the centroid.

2.3 Attention Mechanism

This section mainly introduces two attention mechanisms incorporated in ST-GCN: Channel Attention Module and Spatial Attention Module.

Spatial Attention Module

By incorporating the spatial attention mechanism into the spatio-temporal graph convolutional network, not only can the network parameters be learned, but it can also adaptively capture dynamic relationships between nodes in the spatial dimension. This allows for assigning varying importance to different positions in the human spatiotemporal skeleton graph, enhancing crucial areas and suppressing less significant regions.

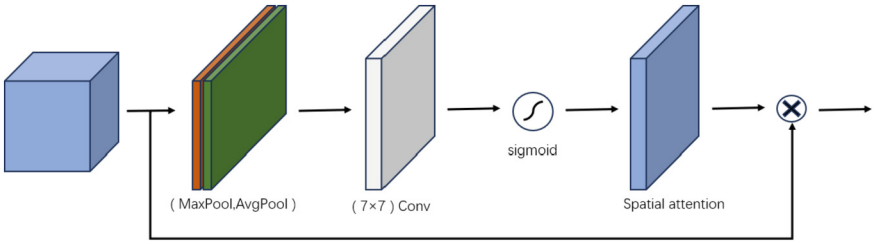


Fig. 3. Spatial attention module.

The spatial attention module, as shown in Fig. 3, is designed as follows in this study. The input features are first subjected to both max pooling and average pooling operations. Then, a convolution operation with kernel size 7×7 is performed, followed by activation through a sigmoid function to generate a feature matrix of size $1 \times H \times W$. Here, H and W represent the height and width of the feature map, respectively.

$$M(F) = [AvgPool(F); MaxPool(F)] \tag{8}$$

$$M_S(F) = \sigma\left(f^{7\times7}(M(F))\right) \tag{9}$$

In this case, F represents the feature map, AvgPool stands for average pooling, and MaxPool refers to maximum pooling. $f^{7\times7}$ represents the convolutional layer with a

kernel size of 7×7. σ denotes the application of the sigmoid activation function, and $M_S(F)$ represents the spatial attention parameter matrix, which is capable of dynamically changing throughout the training process.

Channel Attention Module

After undergoing the spatial attention module, the feature map obtains preliminary spatial features. In order to extract better motion feature representations, the channel attention module is introduced after the spatial attention module. As shown in Fig. 4, the input feature map is transformed from size $C \times H \times W$ to size $C \times 1 \times 1$ through two parallel max pooling layers and average pooling layers. It then passes through a multi-layer perceptron. In this module, the channel dimension of the feature map is first compressed by a factor of $1/r$ (Reduction) through the first fully connected layer. Subsequently, it is expanded back to the original channel dimension via the second fully connected layer. Two activated results are obtained through ReLU activation function. Finally, these two results are element-wise added together, and the output of the channel attention is obtained through a sigmoid activation function. The attention module performs weighted fusion between the obtained feature weights and the corresponding channel feature values of the original convolution in each channel domain. The fused feature map returns to size $C \times H \times W$, allowing different weights to be manifested in the convolutional channel features, thereby extracting key information representing the target. The specific formula is as follows:

$$\begin{aligned} M_c(F) &= \sigma(MLP(AvgPool(F))) + MLP(MaxPool(F)) \\ &= \sigma\left(W_1\left(W_0\left(F^c_{avg}\right)\right) + W_1\left(W_0\left(F^c_{max}\right)\right)\right) \end{aligned} \tag{10}$$

In this context, σ represents the sigmoid activation function, $W_0 \in \mathbb{R}^{C/r \times C}$, $W_1 \in \mathbb{R}^{C \times C/r}$, and r refer to reduction ratios. The variable C represents the size of the feature channel dimension, and $M_c(F)$ denotes the parameter matrix used for channel attention.

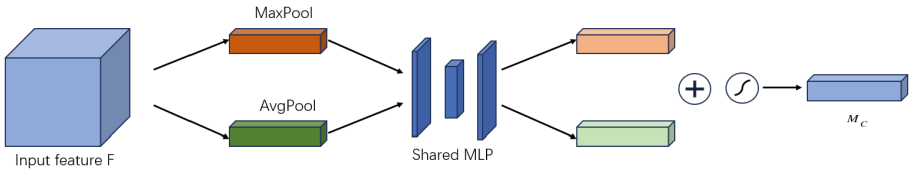


Fig. 4. Channel attention module.

2.4 The Basic Unit and Structure of the Improved ST-GCN

The basic unit of ST-GCN consists of multiple processing modules. Each basic unit includes a spatial graph convolution module, a spatial attention module, a temporal graph convolution module, a channel attention module, and a dropout layer with a dropout rate of 0.5, as shown in Fig. 5. The spatial graph convolution submodule is used to extract motion features from individual frames of the human body. Subsequently, the spatial

attention module assigns corresponding weights to different body parts, continuously updating its parameters during training to help the spatial graph convolution module better extract spatial features from the skeleton graph. The temporal convolution module utilizes a 2D convolution with a kernel size of 9×1 to perform temporal convolution. The resulting features then undergo further feature extraction through the channel attention module. Additionally, both the spatial attention module and the channel attention module are conditioned with a batch normalization layer and an activation function layer. Finally, in order to stabilize the training process, a residual structure is added to the basic unit.

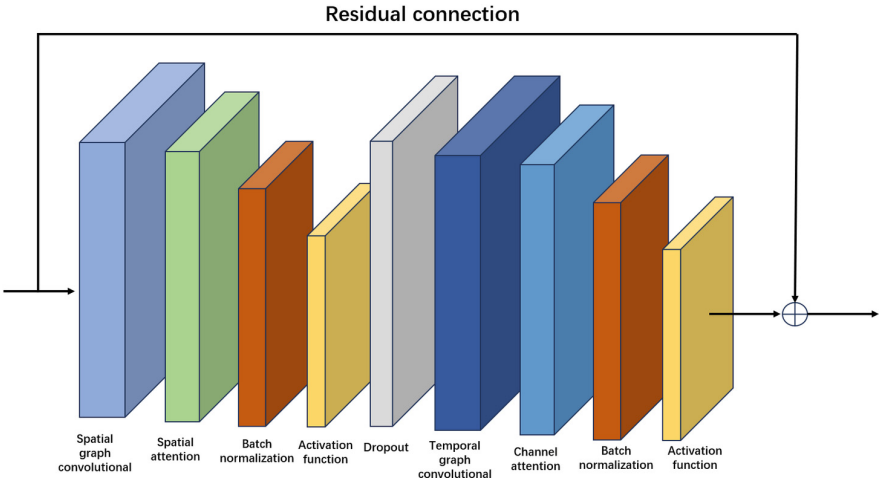


Fig. 5. The basic unit of ST-GCN.

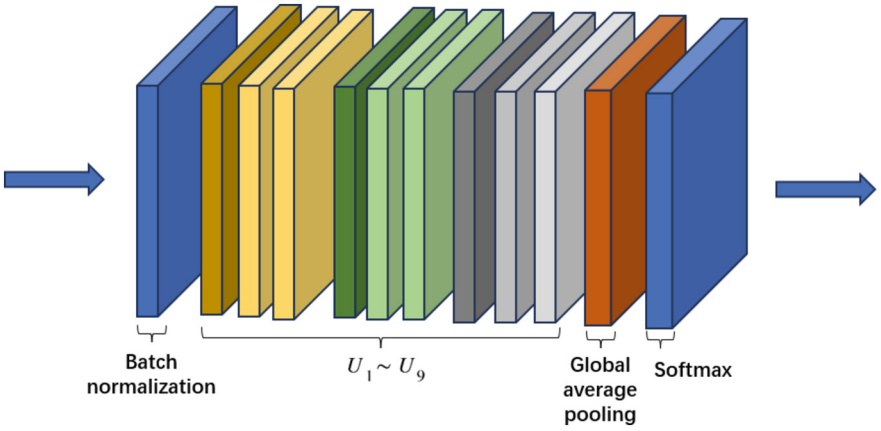


Fig. 6. The overall structure of the network.

The ST-GCN network consists of 9 basic units as described above. It is shown in Fig. 6. Before feeding the skeleton data into the ST-GCN network, it needs to be normalized using a batch normalization module. The input channel is 3 and the output channel is 64. The 9 basic units are named sequentially as a, b with input and output channels both set to 64 and a stride of 1; c with input channel of 64 and output channel of 128, and a stride of 2; d with input channel of 128 and output channel of 256, and a stride of 1. Then, the output of the entire network is passed through global average pooling to obtain a fixed-size feature vector. Finally, this feature vector is fed into a softmax classifier for classification, resulting in the final prediction.

3 Experiments

3.1 Experimental Environment and Dataset

In order to evaluate the performance of the improved spatiotemporal graph convolutional network, experiments were conducted on the NTU RGB + D dataset and a self-built dataset of abnormal human behaviors. All experiments in this study were conducted on a system running Windows 10.0, equipped with an NVIDIA GeForce GXT 1080 graphics card with 32 GB VRAM and an Intel(R) Xeon(R) W-2125 CPU @ 4.00 GHz. The entire deep learning network was implemented using the PyTorch framework, with PyCharm used as the integrated development environment.

1) NTU RGB + D Dataset: The dataset was acquired using three Microsoft Kinect V2.0 sensors, with each camera capturing data from different angles. The collected data includes depth information, 3D skeletal information, RGB images, and infrared sequences. The NTU RGB + D dataset consists of 60 action categories and a total of 56,000 action samples. These samples were performed and recorded by 40 volunteers from different nationalities, age groups, and genders. The dataset authors employed two evaluation criteria: Cross-Subject (CS) and Cross-View (CV). CS refers to training and validation samples that come from different subjects, while CV refers to training and validation samples captured from different camera views.
2) Self-built Abnormal Behavior Dataset: The self-built abnormal behavior dataset comprises three types of actions: falling, fighting, and smoking. These actions were selected to better ensure the safety and civility of public areas such as communities and campuses. As shown in Fig. 7, each action consists of over 100 video sequences, captured from an overhead angle to simulate surveillance cameras in community settings. Each video sequence was clipped to approximately 5 s with a frame rate of 30 frames per second. The videos were processed using the OpenPose algorithm to obtain skeleton data, where each human skeleton consists of 18 nodes. Subsequently, the data was converted into a format compatible with the ST-GCN network for input.

3.2 Comparative Experiments on NTU RGB + D Dataset

In this section, the effectiveness of the proposed improved spatiotemporal graph convolutional network (ST-GCN) was validated on the NTU RGB + D dataset. Based on the

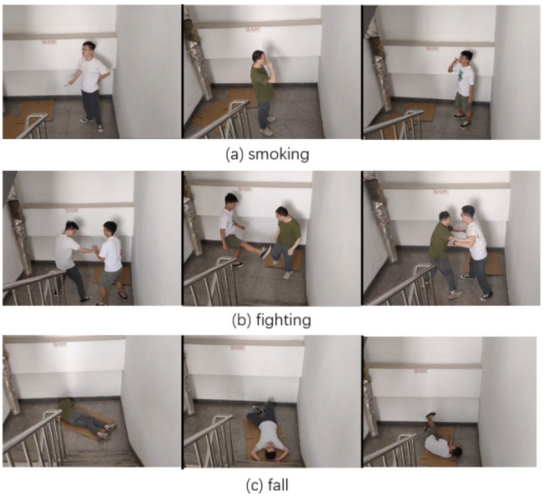


Fig. 7. Self-built dataset example.

characteristics of this dataset, controlled experiments were conducted on two groups of data: cross-subject (CS) and cross-view (CV). To horizontally compare the competitiveness of the proposed algorithm in human skeleton action recognition tasks, advanced algorithms in the current action recognition field were also employed on this dataset for comparison, including the Lie Group method [12] based on manually designed features, CNN + Motion + Trans [13] and TCN [14] based on CNN, ST-LSTM [15] and VA-LSTM [16] based on RNN, and ST-GCN [10], 2s-AGCN [8], and GCN-NAS [19] based on GCN.

For all experiments, the batch size was set to 16, and the training was conducted for 60 epochs. The model utilized the SGD optimizer with a momentum value of 0.9 and weight decay of 0.0001. The initial learning rate was set to 0.1 and was reduced by a factor of 10 at the 20th, 30th, and 40th epoch for continued training. The results obtained are shown in the table below:

According to Table 1, the proposed model achieved better results on this dataset compared to the traditional method of manually designed features, Lie Group, showing significant superiority. In contrast to traditional methods based on Euclidean distance convolution such as TCN and ST-LSTM, the model utilizing human skeletal features effectively reduced background interference, resulting in a significant improvement in recognition accuracy. Compared to the previous version of the ST-GCN network, the inclusion of a dual attention mechanism in our improved model allows for better capture of crucial spatiotemporal features. As a result, we observed a significant improvement in accuracy on the CS and CV datasets, with an increase of 3.1% and 4.8% respectively. These results clearly validate the effectiveness of the improved algorithm. Additionally, when compared to recent advanced GCN networks like 2s-AGCN and GCN-NAS, there are some gaps. However, it is worth noting that these two algorithms introduced dual-stream structure and automatic network search into the GCN network, increasing the

Table 1. Accuracy of different algorithms on the NTU RGB + D dataset.

Method	CS(acc/%)	CV(acc/%)
Lie Group	52.3	79.6
ST-LSTM	70.4	78.8
TCN	74.8	82.6
VA-LSTM	79.5	88.2
ST-GCN	82.3	87.4
CNN + Motion + Trans	84.2	88.5
2s-AGCN	88.2	94.5
GCN-NAS	89.1	95.3
Ours	85.4	92.2

complexity of the networks significantly and adding computational burden and time consumption.

3.3 Comparative Experiments on Self-Built Datasets

This section verified the performance of the improved spatiotemporal graph convolutional network (ST-GCN) on a self-built abnormal behavior dataset. Four experiments were conducted in this section, with one being the experimental group and the other three as control groups. The parameter settings for the four experiments were the same as the ablation experiments on the NTU RGB + D dataset in the previous section.

The first experiment employed the ST-GCN network to process the self-built abnormal behavior dataset to verify the performance of the original ST-GCN model on this dataset. This experiment served as the baseline for the self-built dataset, and the performance of the improved ST-GCN model was compared based on this comparison. The second experiment aimed to validate the effectiveness of the channel attention mechanism on the abnormal behavior dataset. The third experiment aimed to validate the effectiveness of the spatial attention module on the self-built dataset. These two experiments were conducted as control experiments for comparison.

The final experiment was conducted to verify the improved ST-GCN model, which fused the channel attention mechanism and the spatial attention module into the framework of the spatiotemporal graph convolutional network. This model was applied to the abnormal behavior dataset to obtain recognition accuracy.

The experimental results in Table 2 indicated that the model in this study achieved higher accuracy rates on the self-built dataset compared to before the improvements. In terms of horizontal comparison, the recognition accuracy for fighting actions was significantly lower than that of falling and smoking actions under the same algorithm, mainly due to the fact that fighting actions involve multiple targets and are more complex in nature. In terms of vertical comparison, the original ST-GCN algorithm showed better performance on various anomalous behaviors after incorporating both the channel attention mechanism and the spatial attention mechanism. Across the entire dataset, the

Table 2. The accuracy rates of each method on self-built datasets.

method	Fall (acc/%)	Fight (acc/%)	Smoking (acc/%)	All (acc/%)
ST-GCN	88.6	85.3	88.4	87.3
CA + ST-GCN	89.9	85.8	90.1	89.6
SA + ST-GCN	90.9	86.1	90.3	90.1
CA + SA + ST-GCN	93.8	89.4	92.5	92.7

improved algorithm demonstrated a 5.4%increase in accuracy compared to before the improvements, indicating significant improvement.

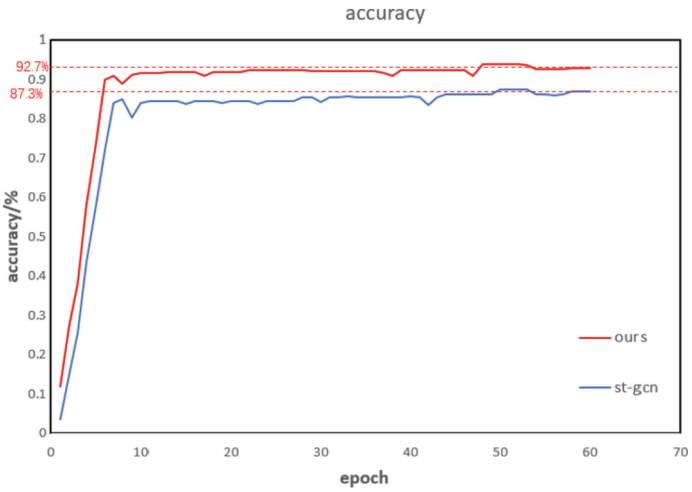


Fig. 8. Comparison chart of accuracy improvement before and after algorithmic recognition.

The figure in Fig. 8 compares the accuracy of the algorithm before and after the improvement. Accuracy is defined as the ratio of the number of correct model predictions to the total number of samples. In the ST-GCN algorithm, the accuracy steadily increases from rounds 0 to 10, stabilizes during subsequent training, and eventually reaches an accuracy of 87.3%. The improved algorithm also exhibits rapid improvement in accuracy from rounds 0 to 10, stabilizes during further training, and ultimately achieves an accuracy of 92.7%. This represents a 5.4% increase compared to the pre-improvement accuracy, validating the conclusion that the dual attention mechanism enhances the performance of the model.

In Fig. 9, the comparison of loss functions before and after the improvement is presented. The loss function is a crucial component in deep learning, determining the robustness of the model. In the ST-GCN algorithm, the loss function rapidly decreases from rounds 0 to 20 and converges to 0 around round 40. In contrast, the improved

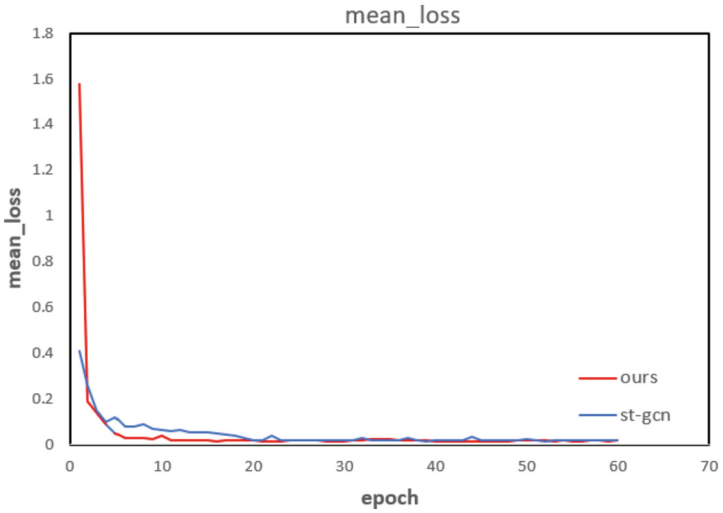


Fig. 9. Comparison chart of loss function curves before and after algorithm improvement.

algorithm incorporates a gradually decaying learning rate, which facilitates a faster decrease in the loss function. It converges to 0 around round 20, optimizing the overall performance of the algorithm. Clearly, based on the experimental data, the proposed algorithm strengthens the stability of the model.

4 Conclusion

This paper proposed a dual-attention spatiotemporal graph convolutional neural network and applied it to skeleton-based anomaly behavior detection. By adding a spatial attention mechanism after the spatial convolutional layer of the existing ST-GCN algorithm, the network focused more on joint information in the human skeleton data that had a greater impact on actions. Simultaneously, a channel attention mechanism was introduced into the temporal convolutional module to obtain channel features that were more important for output channels, further improving the model's performance. Furthermore, the proposed algorithm was compared with other algorithms on the widely recognized NTU RGB + D dataset, surpassing traditional handcrafted features and some commonly used deep learning networks in terms of accuracy, a critical performance metric. Finally, the algorithm was deployed for anomalous behavior detection, and on a self-collected anomaly behavior dataset, it demonstrated improved performance compared to the original algorithm. The recognition accuracy reached 92.7%, meeting the security requirements of certain communities and schools.

Acknowledgments. This work was supported by the National Key Research and Development Project (2019YFB2101902).

References

1. Cai, Q., Deng, Y., Li, H., et al.: Review of human behavior recognition methods based on deep learning. Comput. Sci. **47**(4), 85–93 (2020)
2. Fernando, B., Gavves, E., Oramas, J.M., et al. Modeling video evolution for action recognition, In: Conference on Computer Vision and Pattern Recognition,pp. 5378–5387. IEEE, Boston (2015)
3. Huang, S.: Research on Human Action Recognition Based on Skeleton. Shanghai Shanghai Jiaotong University (2014). (in Chinese)
4. Guan, S., Zhang, Y.: 3D human behavior recognition based on convolution network of residual spatiotemporal graph. Comput. Appl. Softw. **37**(3), 198–201 (2020)
5. Wan, X.: Research on 3D Human Behavior Recognition Based on Spatiotemporal Structure Relationship, pp. 1–3. Suzhou University, Suzhou (2018). (in Chinese)
6. Ding, X., Yang, K., Chen, W.: A semantics-guided graph convolutional network for skeleton-based action recognition. In: IEEE. IEEE International Conference on Innovation in Artificial Intelligence, pp. 130–136. IEEE, Xiamen (2020)
7. Shi, L., Zhang, Y.F., Cheng, J., et al.: Action recognition via pose-based graph convolutional networks with intermediate dense supervision. Pattern Recogn. 121, 108170 (2022)
8. 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), pp. 12026-12035. Long Beach, CA, USA (2019)
9. Velikovi, P., Cucurull, G., Casanova, A., et al.: Graph attention networks. https://arxiv.org/pdf/1710.10903.pdf. Accessed 06 Sept 2020
10. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: AAAI,2018 AAAI. The National Conference on Artificial Intelligence, pp. 7444–7452. San Francisco (2018)
11. Xu, B., Huang, J., et al.: Review of graph convolutional neural networks. J. Comput. Sci. Technol. **43**(05), 755–780 (2020)
12. Vemulapalli, R., Arrate, F., Chellappa, R.: Human action recognition by representing 3D human skeletons as points in a lie group. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 588–595. IEEE, Columbus (2014)
13. Li, C., Zhong, Q., Xie, D., et al.: Skeleton-based action recognition with convolutional neural networks. In: 2017 IEEE International Conference on Multimedia & Expo Workshops, pp. 597–600. IEEE, Hong Kong(2017)
14. Kimts, T., Reitera, A.: Interpre table 3D human action analysis with temporal convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1623–1631. IEEE, Honolulu (2017)
15. Liu, J., Shahroudy, A., Xu, D., Wang, G.: Spatio-temporal LSTM with trust gates for 3D human action recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 816–833. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_50
16. Zhang, P., Lan, C., Xing, J., et al.: View adaptive neural networks for high performance skeleton-based human action recognition. In: IEEE International Conference on Computer Vision, pp. 2136–2145. IEEE, Venice (2017)
17. Peng, W., Hong, X., Zhao, G.: Learning graph convolutional network for skeleton-based human action recognition by neural searching. In: The Thirty-Fourth AAAI Conference. Advancement of Artificial Intelligence (AAAI), vol. 34, no. 03, pp. 2669–2676. New York (2020)