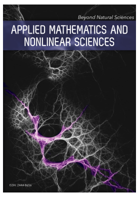
Applied Mathematics and Nonlinear Sciences 7(2) (2022) 703–712



Applied Mathematics and Nonlinear Sciences

https://www.sciendo.com

Spatial–temporal graph neural network based on node attention

Qiang Li[1], Jun Wan[2†], Wucong Zhang[2], Qian Long Kweh[3]

[1] School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China

[2] Midea Intelligent Life Research Institute, Midea Real Estate Holding Limited, Foshan, China

[3] Canadian University Dubai

Submission Info

Communicated by Juan Luis García Guirao
Received May 23rd 2021
Accepted September 27th 2021
Available online April 8th 2022

Abstract

Recently, the method of using graph neural network based on skeletons for action recognition has become more and more popular, due to the fact that a skeleton can carry very intuitive and rich action information, without being affected by background, light and other factors. The spatial–temporal graph convolutional neural network (ST-GCN) is a dynamic skeleton model that automatically learns spatial–temporal model from data, which not only has stronger expression ability, but also has stronger generalisation ability, showing remarkable results on public data sets. However, the ST-GCN network directly learns the information of adjacent nodes (local information), and is insufficient in learning the relations of non-adjacent nodes (global information), such as clapping action that requires learning the related information of non-adjacent nodes. Therefore, this paper proposes an ST-GCN based on node attention (NA-STGCN), so as to solve the problem of insufficient global information in ST-GCN by introducing node attention module to explicitly model the interdependence between global nodes. The experimental results on the NTU-RGB+D set show that the node attention module can effectively improve the accuracy and feature representation ability of the existing algorithms, and obviously improve the recognition effect of the actions that need global information.

Keywords: Action recognition, skeletons, spatial–temporal graph convolution, attention mechanism

1 Introduction

The action recognition technology has been widely used in video understanding, human–computer interaction, intelligent control and other fields, but restricted by background, illumination, occlusion and camera jitter. Thus, the accuracy of action recognition algorithm still faces a great challenge.

† Corresponding author.
Email address: skysweetgrape@163.com

ISSN 2444-8656

doi:10.2478/amns.2022.1.00005

In the early stage of video understanding field, the approaches based on manual feature representation were the main research direction. Dense trajectories (DT) [1] and its improved version – improved Dense trajectories (iDT), show the best performance among the methods based on manual feature representation. With the development and maturity of deep learning technology, researchers turned to use deep learning algorithms for action recognition: Simonyan [3] proposed the double-stream method based on RGB and optical flow for action recognition; Feichtenhofer [4] introduced residual structure into double-stream convolutional network for information exchange; Tran proposed C3D [5] and its improved C3D [6] structure that used 3D convolution to learn static appearance and action characteristics. However, the above video feature-based methods (two-stream method, 3D convolution method) tend to be affected by background illumination, camera movement and others, cannot represent the human action sequence information very well, and has unobvious data concentration performance on complex actions.

Benefitted from the improved performance of human pose estimation and other algorithms, the method based on skeletons is not affected by background, illumination and other factors, and is more and more popular in the action recognition field. The traditional skeleton point method [7] requires the establishment of manual features and traversal rules, which is inefficient; the common skeleton method based on deep learning is to construct skeletons information into coordinate vector or pseudo image and input them into CNN or recurrent neural network (RNN) for action recognition [8–12]; the graph convolution methods [13–15] by constructing human skeletons points as the graph nodes, and the connection information between skeletons as graph edges use the method similar to the traditional 2D convolution method in the skeleton graph for action recognition, achieving significant results.

The spatial–temporal graph convolutional neural network (ST-GCN) method modelled the dynamic skeleton [13] based on the time sequence representation of human joint position, and extended the graph convolution into a spatial–temporal graph convolutional network. As the first method using graph convolution neural network for skeleton-based action recognition, it is different from the previous methods, because it can implicitly learn the human body information of various body parts by using the locality and time dynamics of graph convolution. By eliminating the requirement of manual allocation of various of human body parts, the model can be designed easier and can effectively learn better action representation. However, the convolution operation in the ST-GCN is performed only on the 1-neighbour of the root node, so the modelling and representation on the global node information cannot be realised. For example, the interactive joints for brushing teeth, clapping and other actions are not in the adjacent position, so it is necessary to learn the relationship of these joints through the attention mechanism to improve the action recognition performance. The common attention methods include SENet [16], convolutional block attention module (CBAM) [17] and non-local network [18].

To solve the above problems, this paper proposed an ST-GCN algorithm based on node attention (NA-STGCN). In the NA-STGCN network, we introduced the attention module to help the network focus on the connection relationship between different nodes (including adjacent and non-adjacent nodes) and learn the importance of nodes. Specifically, we introduced the attention mechanism of SENet into the convolutional layer of ST-GCN to learn the correlation between nodes. The effect of introducing attention module to network was verified by experiments. The experimental results on NTU-RGB+D show that our NA-STGCN with node attention module introduced has improved accuracy over ST-GCN.

The second part of this paper introduces the related work, and the third part introduces the original ST-GCN model and our proposed NA-STGCN model; the fourth part is the experimental results and analysis, and the last part is the algorithm summary.

2 Related work

With the rapid development of human pose estimation and graph neural network, now most common action recognition methods which are based on skeleton can be categorised into three methods: CNN, RNN and graph convolutional network-based methods.

The traditional method in [7] requires traversal rules and manual features to realise the skeleton action recognition, which is inefficient and inaccurate. Recently, deep learning has achieve great success which makes the deep learning based skeleton modelling methods rather hot now. As for CNN based methods, Liu et al. [8] put forward a new type of two-stream model which uses the 3D CNN. The model is very innovative and nobody has proposed it before this; Li et al. [9] put forward a new scheme called the global spatial aggregation scheme. The new scheme is better than local aggregation in the point of joint co-occurrence features. As for RNN-based methods, according to the principle of RNN with the long short-term memory (LSTM) and the convolutional neural network; Zhang et al. [10] designed VA-RNN and VA-CNN view adaptive neural networks. In order to analyse the hidden sources of information which has something to do with action, within the input data over the two domians concurrently, Liu et al. [11] tried to extend RNN-based methods to spatio-temporal domains; According to the physical structure of humans, Du et al. [12] divided the human skeleton into five different parts, and then separately fed the five different parts to five bidirectional recurrently. Based on skeleton action recognition, the method in [12] is an end-to-end hierarchircal RNN. Yan et al. [13] put forward the ST-GCN. The new dynamic skeletons model can automatically learning both the spatial and temporal patterns from data. Therefore, it is superior to the previous methods and can break out of limitations. In addition to enhancing expressive ability, the data patterns can also improve the generalisation ability. In graph convolution operation, Shi et al. [14] used the method of non-local attention to model the multi-level semantic information. In this way, the flexibility of graph construction model is increased; what is more, the generality to adapt to sundry data samples is also increased. Inspired by deformable part-based models (DPMs), Thakkar et al. [15] designed a part-based graph convolutional network which improves the recognition performance when compare with a model using the entire skeleton graph. In the network, the skeleton graph is divided into four subgraphs and shared joints between joints.

3 Method

3.1 Original ST-GCN model

In the method proposed in [13], the CUHK team put forward an idea to extend the graph neural network to a spatial–temporal graph model, which is also called ST-GCN, to design a general representation of skeleton sequences for action recognition. This is shown in Figure 1(a). The new model is built based on the sequence of skeleton graph, where each node corresponds to a joint of the human body. There are two different edge types, one is the space edge consistent with the natural connectivity of the joint, and the other is the time edge connected to the same joint on a continuous time step. A number of spatial–temporal graph convolution layers

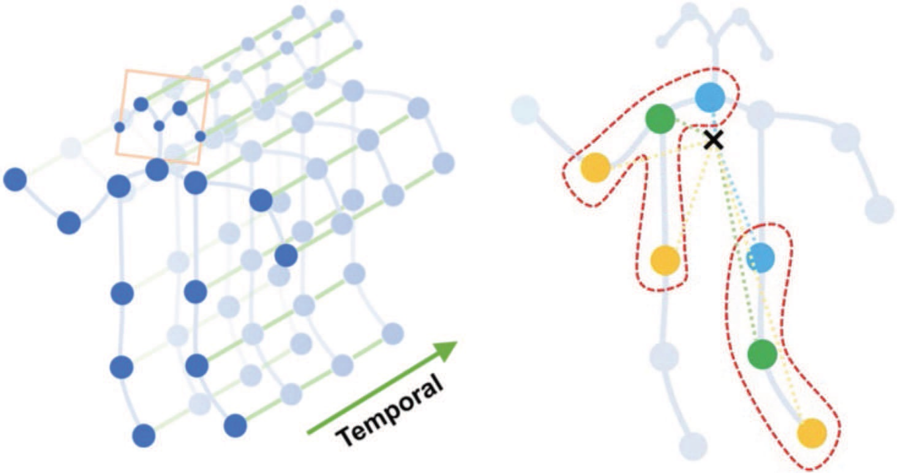


Fig. 1 (a) Spatial–temporal skeletal graph. (b) Partitioning strategy graph.

are constructed to extract feature graph information, and then the SoftMax classifier is often used to predict.

According to the study on motion analysis, the space structure of the graph in ST-GCN is divided as shown in Figure 1(b). Each node of partition 1 is divided into three subsets. Taking shoulder nodes as an example, the first subset is the node itself (green point), the second subset is the adjacent node sets (blue point) closer to the whole skeleton centre of gravity, and the third subset is the neighbouring node sets (yellow points) further away from the centre of gravity. Each colour represents a learnable weight for learning the information between nodes.

In a single frame, the graph convolution process of ST-GCN can be expressed by the following equation:

$$f_{out} = \sum_{j} \left(\Lambda_j^{-\frac{1}{2}} A_j \Lambda_j^{-\frac{1}{2}} \right) \otimes M_j f_{in} W_j \tag{1}$$

where f_{in} represents the input feature graph of $C_{in} \times T \times V$, f_{out} represents the input feature graph of $C_{out} \times T \times V$, C_{in} and C_{out} represent the number of input and output channels, T represents the number of video frames, V represents the number of skeletons, A represents the adjacency matrix of $18 \times 18 \times 3$, 18 is the number of skeleton points in the human body (taking openpose as example), 3 is the number of partitions in partition strategy, the element A_{ij} in matrix A represents the connectivity from node i to node j; $A_0 = I$ represents the node's own connection matrix, A_1 represents the near-centre subset, A_2 represents the far-centre subset, $\Lambda_j^{ii} = \sum_k \left(A_j^{ki} \right) + \alpha$ represents a normalised diagonal matrix; set α to 0.001, in order to avoid the matrix A from being null; $\otimes$ represents the point multiplication between matrices; W_j is the convolution kernel of $C_{out} \times C_{in} \times 1 \times 1$, used for 1×1 convolution operation; and M is the learnable attention module of $V \times V$, used for learning the importance of different skeletal points, with an initial value of all 1. It can be seen from Eq. (1) that when the element in the matrix A is 0, there is no connection between nodes; no matter what the M value is, the output result is still 0, so M can only learn the nodes in one-neighbour node.

3.2 Our proposed NA-STGCN model

In the ST-GCN network, the receptive field of convolution kernel is only in the range of one neighbour, so it can only extract the local feature information. Global feature information plays a more important role in actions where the adjacent distance between nodes is >1, such as clapping, drinking water, and so on. Therefore, we proposed a node attention model called NA-STGCN, which integrates the attention method of SENet into the ST-GCN. By introducing the attention module, the network can focus on the connection between different nodes (including adjacent and non-adjacent nodes) and learn the importance of nodes.

Figure 2 is a schematic diagram of the NA-STGCN network structure. The GCN represents the spatial convolution operation and the TCN represents the temporal convolution operation. The left part of Figure 2 is the overall structure diagram. There are nine GCN + TCN modules (layer 1-layer 9) in the network, and the residual part is added to each layer. The number of output channels of layer 1-layer 3 is 64, of layer 4-layer 6 is 128 and of layer 7-layer 9 is 256.

The idea of attention module algorithm in our study comes from SENet. In SENet, the SE module first executes compress operation to the feature graph which is obtained by convolution, to get the global feature at channel level, then it executes excitation operation to the global feature, in order to learn the relationship and get weights of different channels; Therefore, it multiplies by the original feature graph to get the final feature. Virtually, on the channel dimension, the SE module executes the attention or gating operation. The attention mechanism has the advantage to focus on the channel features which has the most information and it can also suppress unimportant channel characteristics at the same time.

However, our proposed attention module is a little different from SENet. As shown in the dotted line box on the right side of Figure 2, we first globally pool the $c \times t$ dimensions of the GCN output feature graph to obtain the node-level features, and then carry out excitation operation to the node features to learn the relationship between nodes because the number of nodes is small, for example, 18 or 25, then carry out squeeze operation to

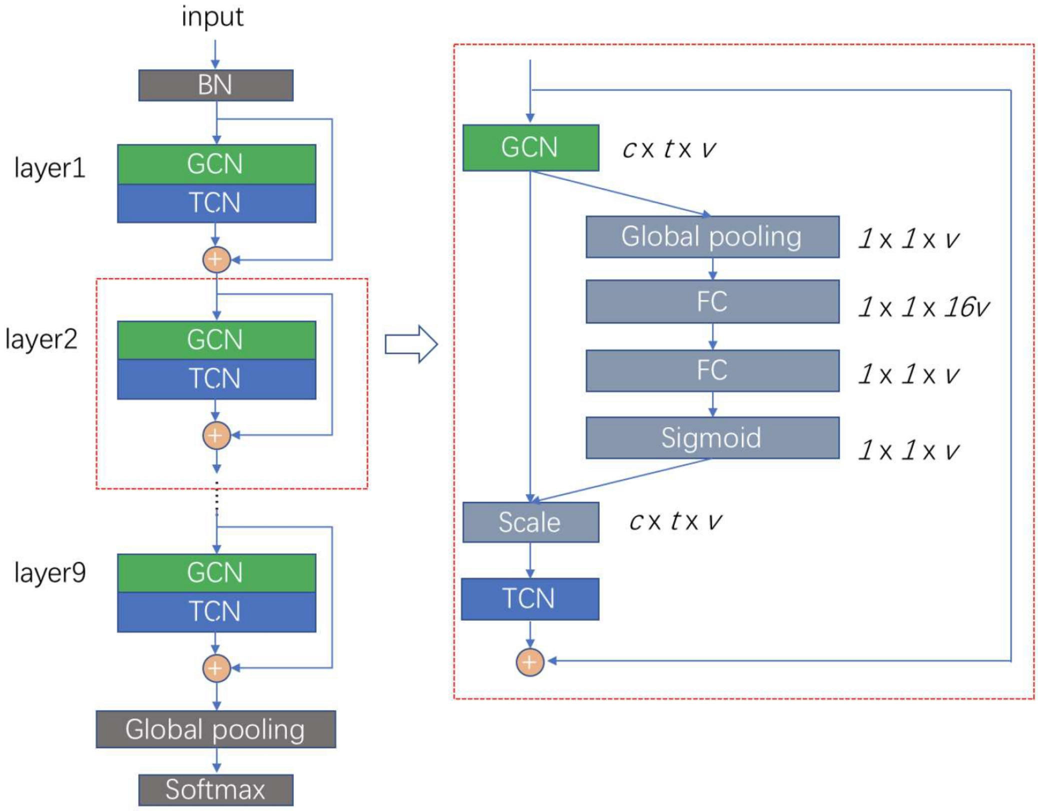


Fig. 2 Structure diagram of NA-STGCN. GCN, graph convolutional neural network; NA-STGCN, node attention spatial–temporal graph convolutional neural network.

obtain the importance of different nodes; finally, multiply by the original feature graph to get the final feature. By adding such an attention learning module, NA-STGCN learned the correlation between different nodes. In Section 4, according to the conclusion in [19] that used non-local attention in ST-GCN and added the attention modules in layer 2 and layer 3 acheive a better result, we experimentally explored the performance effect of adding our node-attention modules in layers 2 and 3. Moreover, in order to analyse the importance of every node, we estimate the class activation map (CAM [20]) of every node. Finally, we compare NA-STGCN with previous representative methods.

4 Experiment

To verify the extract ability of global features and recognition performance of NA-STGCN structure, the study on skeleton behaviour recognition is carried out based on NTU-RGB+D [21]. The experimental platform is as follows: Linux system, i7-7700 CPU and 1070 graphics card, 16 GB memory, and Pytorch depth learning framework.

4.1 Data sets

NTU-RGB+D data set is a public data set marked with 3D node information, which is used for identifying human actions. It contains 56,880 action fragments and 60 action categories. All the action clips were completed by 40 volunteers in a laboratory environment, and photographed from three cameras of the same height but different levels: −45°, 0°, 45°. The data set uses the 3D joint positions detected by Kinetic sensor in each frame. Each experimenter has 25 joints in the skeletal sequence. There are two partition methods for NTU-RGB+D data set, one is cross-action object partition (CS) and the other is cross-view partition (CV). Both are used to test the recognition accuracy of the model.

4.2 Experimental allocation

The experiments are carried out based on PyTorch deep learning framework; the optimisation strategy uses stochastic gradient descent (SGD); the momentum of Nesterov is 0.9; the initial learning rate is set to 0.1; and the learning rate decay is set (for 10 and 50 rounds learning rate, decay is 0.01 and 0.001); the training batch size is 16; the cross entropy is used as the loss function of gradient backpropagation; and the weight attenuation factor is 10^{-4}.

4.3 Experimental results and analysis

4.3.1 Loss comparison between NA-STGCN and ST-GCN

In order to verify that the NA-STGCN has better global information modelling ability compared with the traditional spatial–temporal graph convolution, the loss comparison experiments were carried out on NTU-RGB+D data sets in CV partition. Figure 3 shows the loss curves of ST-GCN and NA-STGCN changing with the number of training epoches. As shown in Figure 3, Na-STGCN has a faster convergence rate than ST-GCN, and its loss value is also lower.

4.3.2 Comparison of node activation response maps between NA-STGCN and ST-GCN

In order to verify NA-STGCN has the ability of global information modelling on nodes and the ability of learning the importance between different nodes, we estimated the response values of different nodes of people in a specific action segment by using the method in [20]. Figure 4 shows the node response maps of NA-STGCN and ST-GCN in the actions of clapping and brushing. The reason that we chose these two actions as the analysis examples is that these actions focus more on the information exchange between non-adjacent joints.

According to Figure 4(a) and 4(b), in terms of the clapping action, compared with ST-GCN, NA-STGCN has larger response values at the hand nodes, elbow nodes and shoulder nodes, and smaller response values at the trunk and lower limbs nodes, indicating that through the node attention method, NA-STGCN has learned the

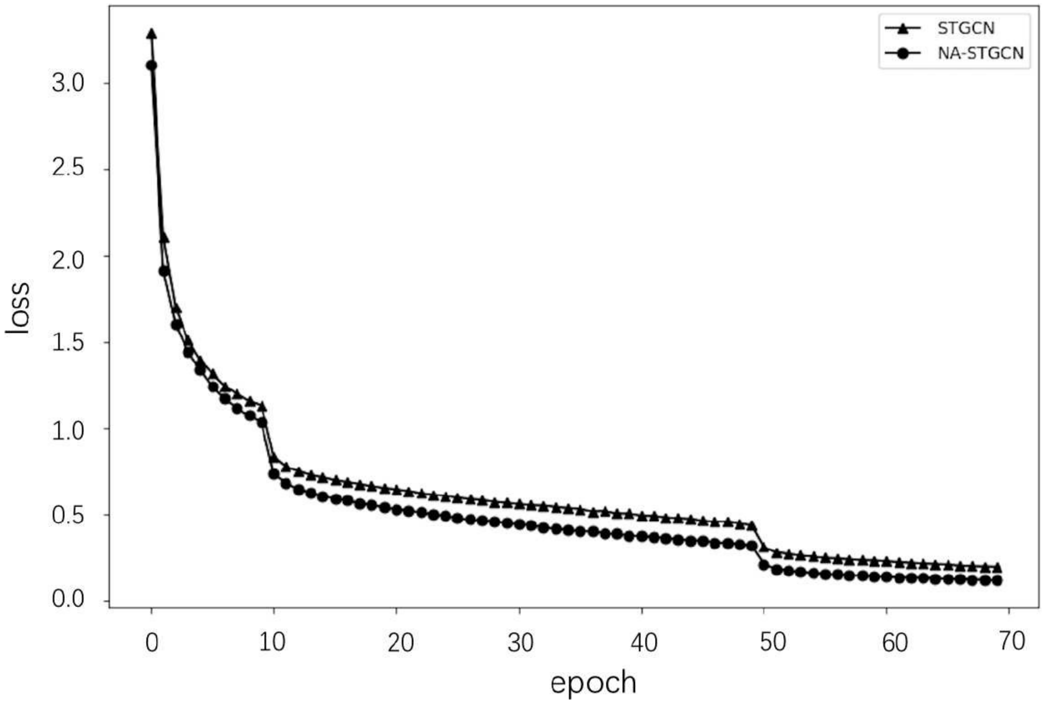


Fig. 3 Change curve of loss values. NA-STGCN, node attention spatial–temporal graph convolutional neural network; ST-GCN, spatial–temporal graph convolutional neural network.

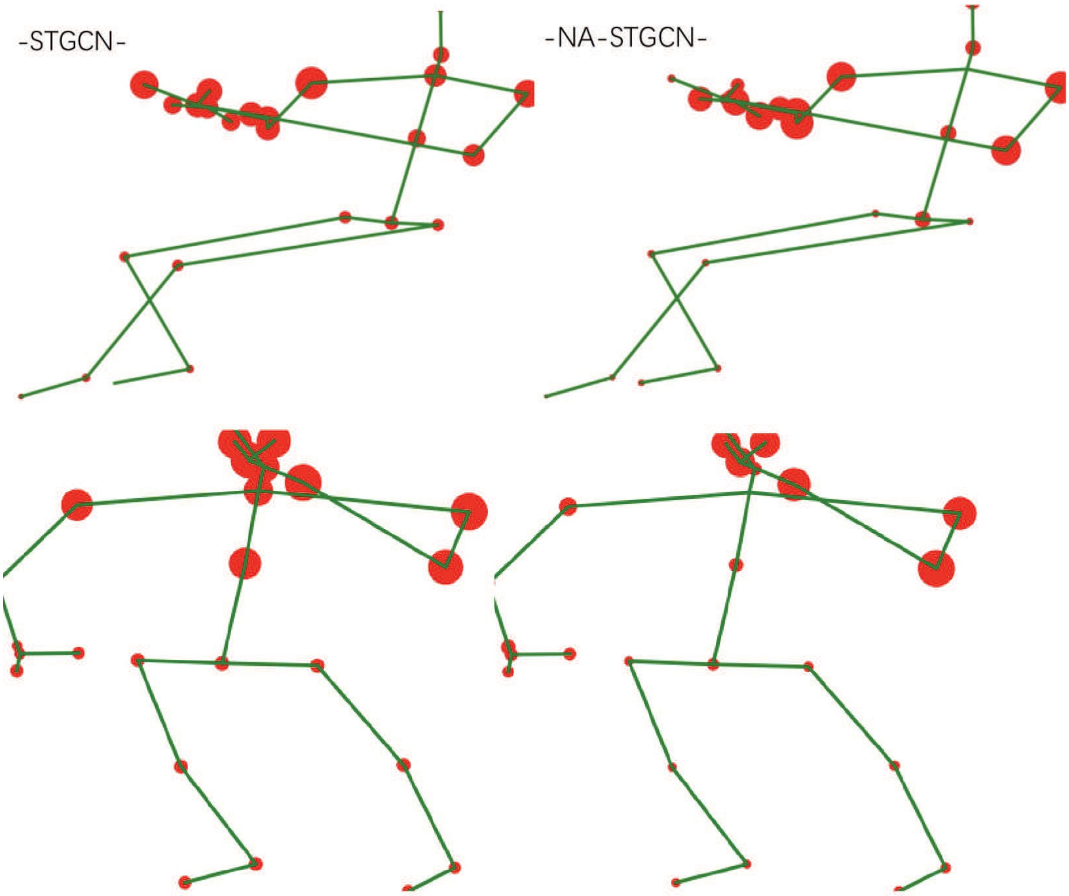


Fig. 4 Node activation response maps: (a) top left: ST-GCN clapping action; (b) top right: NA-STGCN clapping action; (c) bottom left: ST-GCN brushing action; (d) bottom right: NA-STGCN brushing action. NA-STGCN, node attention spatial–temporal graph convolutional neural network; ST-GCN, spatial–temporal graph convolutional neural network.

importance related to the category of actions for different nodes. Similarly, it can be seen from Figure 4(c) and 4(d) that, in terms of the brushing action, compared with ST-GCN, the NA-STGCN has larger response value in the left hand joint and neck nodes than in other parts. Since this action is a person brushing with the left hand, the response result of NA-STGCN is more in line with the actual situation.

Overall, NA-STGCN can model the global information of nodes through node attention method, and adaptively learn the importance of different nodes.

4.3.3 Comparison of results between NA-STGCN and representative methods

In order to verify the recognition accuracy of the NA-STGCN model, experiments and tests were carried out on the NTU-RGB+D skeleton action data sets in CV and CS partition modes. The experimental results are shown in Table 1.

As can be seen from Table 1, in the NTU-RGB+D data sets in CS and CV partition modes, the recognition accuracy of NA-STGCN is 85.8% and 89.3%, respectively. Compared with ST-GCN, the recognition accuracy of NA-STGCN in the CS and CV partition modes is improved by 4.3% and 1.0%, respectively. It can be seen that the accuracy of NA-STGCN in the CS partition mode is significantly improved, but the recognition accuracy in the CV partition mode is limited, at only 1%, which may be because the camera angle affects the learning of node attention.

In conclusion, according to the experimental results of (1)–(3), NA-STGCN shows higher accuracy and

Table 1 Comparison with representative methods (%).

Model	CS	CV
Two-Stream 3DCNN [8]	66.8	72.6
TCN [22]	74.3	83.1
Clip + CNN + MTLN [23]	79.6	84.8
VA-LSTM [10]	79.4	87.6
ST-GCN [13]	81.5	88.3
NA-STGCN (ours)	85.8	89.3

CS, cross-action object partition; CV, cross-view partition; LSTM, long short-term memory; NA-STGCN, node attention spatial–temporal graph convolutional neural network; ST-GCN, spatial–temporal graph convolutional neural network.

faster convergence speed compared with ST-GCN and other representative methods. Moreover, the study in 2 proves that the NA-STGCN can model the global information of nodes and adaptively learn the importance of different nodes.

5 Conclusion

In this study, a new skeleton graph neural network is proposed, namely NA-STGCN solving the algorithm flaw (only learns information from one-neighbour node) of the original ST-GCN model. Specifically, we add the SENet attention mechanism to the GCN layer, to enable the network to learn the interactive information of all joints. Therefore, the NA-STGCN can learn long-range relationships in skeletal action sequences, which overcomes the defect in ST-GCN which can only learn the information of one-neighbour node. The experiments carried out on NTU-RGC+D show that the recognition accuracy of NA-STGCN is better than that of ST-GCN. Moreover, in order to analyse the importance of every node, we estimate CAM of every node and the results show that NA-STGCN can focus on the connection relationship between different nodes (including adjacent and non-adjacent nodes) and learn the importance of nodes. In the future, we will study the effects of multiple attention mechanisms on ST-GCN, and explore the method of improving ST-GCN accuracy by combining multi-modality.

References

[1] Wang, H., Kläser, A., Schmid, C., et al.: ‘Dense trajectories and motion boundary descriptors for action recognition’, International Journal of Computer Vision, 2013, 103, (1), pp. 60–79.

[2] Wang, H., Schmid, C.: ‘Action recognition with improved trajectories’. International Conference on Computer Vision (ICCV), Sydney, NSW, Australia, October 2013, pp. 3551–3558.

[3] Simonyan, K., Zisserman, A.: ‘Two-stream convolutional networks for action recognition in videos’. Neural Information Processing Systems (NIPS), Montreal, Canada, December 2014, pp. 2136–2145.

[4] Feichtenhofer, C., Pinz, A., Wildes, R. P.: ‘Spatiotemporal residual networks for video action recognition’. Neural Information Processing Systems (NIPS), Barcelona, SPAIN, December 2016, pp. 3476–3484.

[5] Tran, D., Bourdev, L., Fergus, R., et al.: ‘Learning spatiotemporal features with 3D convolutional networks’. International Conference on Computer Vision (ICCV), Santiago, Chile, December 2015, pp. 4489–4497.

[6] He, K., Zhang, X., Ren, S., et al.: ‘Deep residual learning for image recognition’. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, Nevada, USA, June 2016, pp. 770–778.

[7] R. Vemulapalli, F. Arrate, and R. Chellappa, “Human action recognition by representing 3d skeletons as points in a lie group,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 588–595, IEEE, Columbus, OH (2014).

[8] Liu H, Tu J, Liu M. Two-stream 3D convolutional neural network for skeleton-based action recognition [J/OL]. [2017-03-23].

[9] Li C, Zhong Q, Xie D, et al. Co-occurrence Feature Learning from Skeleton Data for Action Recognition and Detection

with Hierarchical Aggregation [C]. Twenty-Seventh International Joint Conference on Artificial Intelligence {IJCAI-18. 2018.

[10] Zhang P, Lan C, Xing J, et al. View Adaptive Neural Networks for High Performance Skeleton-Based Human Action Recognition [J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2019.

[11] Liu J, Shahroudy A, Dong X, et al. Spatio-Temporal LSTM with Trust Gates for 3D Human Action Recognition[J]. 2016.

[12] Du, Y.; Wang, W.; and Wang, L. 2015. Hierarchical recurrent neural network for skeleton based action recognition. In CVPR, 1110–1118.

[13] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in Thirty-Second AAAI Conference on Artificial Intelligence, pp. 7444–7452, AAAI Press, New Orleans, Louisiana, USA (2018).

[14] L. Shi et al., "Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 12026–12035 (2019).

[15] K. Thakkar, P J. Narayanan, "Part-based Graph Convolutional Network for Action Recognition," arXiv preprint arXiv:1809.04983, 2018.

[16] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7132–7141, IEEE, Salt Lake City, UT (2018).

[17] S. Woo et al., "CBAM: Convolutional Block Attention Module," in Proceedings of the European Conference on Computer Vision (ECCV), Lecture Notes in Computer Science, vol 11211, pp. 3–19, Springer, Cham (2018).

[18] X. Wang et al., "Non-local neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794-7803, IEEE, Salt Lake City, UT, USA (2018).

[19] Kong, Y., Li, L., Zhang, K., Ni, Q., & Han, J. (2019). Attention module-based spatial–temporal graph convolutional networks for skeleton-based action recognition. Journal of Electronic Imaging, 28(4), 1.

[20] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning Deep Features for Discriminative Localization. CVPR. IEEE Computer Society.

[21] Shahroudy A, Liu J, Ng T T, et al. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis [J]. IEEE Computer Society, 2016:1010–1019.

[22] Tae S K, Austin R. Interpretable 3D human action analysis with temporal convolutional networks [C]. Proc of IEEE Computer Vision and Pattern Recognition Workshops. New York: IEEE, 2017: 1623–1631.

[23] Oord A V D, Dieleman S, Zen H, et al. Wavenet: a generative model for raw audio [J/OL]. [2016-09-12].