ICIC Express Letters
Part B: Applications
Volume **15**, Number **2**, February **2024**

ICIC International ©2024 ISSN 2185-2766
pp. **171–178**

ATTENTIONAL MULTI-SCALE UNIFIED SPATIOTEMPORAL GRAPH CONVOLUTIONAL NETWORK FOR SKELETON-BASED ACTION RECOGNITION

Xianshan Li[1,3], Jingwen Kang[1,3] and Fengda Zhao[1,2,3,*]

[1]School of Information Science and Engineering
[3]Key Laboratory for Software Engineering of Hebei Province
Yanshan University
No. 438, West Hebei Avenue, Qinhuangdao 066004, P. R. China
xjlxs@ysu.edu.cn; kjw@stumail.ysu.edu.cn
*Corresponding author: zfd@ysu.edu.cn

[2]School of Information Science and Engineering
Xinjiang University of Science and Technology
No. 360, Jinhe Road, Korla Economic and Technology Development Zone, Korla 841000, P. R. China

Received June 2023; accepted August 2023

Abstract. *Skeleton-based human action recognition has become a hot research topic in computer vision recently. However, the skeleton action recognition method based on graph convolution has the problems that spatiotemporal separated graph convolution cannot well establish multi-order cross-spatiotemporal connections of human joint points. In this paper, we propose an attentional multi-scale unified spatiotemporal graph convolutional network. The unified spatiotemporal graph convolution directly constructs the cross-spatiotemporal human skeleton connection, realizing barrier-free spatiotemporal communication. The multi-scale graph convolution ensures the effectiveness of long-distance features at different step sizes. Moreover, an external attention mechanism is added to extract potential connections between different samples. In addition, according to the physical connection between the skeleton joints, a six-stream architecture is designed to further strengthen the expressiveness of the model. Experiment results on two skeleton datasets, NTU RGB+D 60 and NTU RGB+D 120, show that this model reaches an advanced level compared to other methods.*

Keywords: Action recognition, Skeleton data, Graph convolution, Attention module

1. **Introduction.** Human action recognition classifies human behavior by analyzing human body posture and movement trajectory. At present, human action recognition technology has been widely used in the fields of video surveillance and virtual reality [1] and has important research value and practical significance. Human 3D skeleton data are relatively insensitive to background changes and can provide more accurate human body posture information in 3D space. Compared with traditional RGB video and depth images, skeleton data have more advantages and broader application prospects.

Most of the previous studies [2-10] on human skeleton action recognition based on the graph convolution method divide the skeleton information into two dimensions of space and time to perform graph convolution operations separately. However, this approach hinders the direct cross-spatiotemporal connections of human joints. At the same time, how to establish the connection between the nodes and their neighboring nodes of different steps is also a problem worth considering. Meanwhile, the current method [3-5] prefers to explore deeper connections between skeleton joints and ignores correlations among action samples. In addition, most of the networks [3,4] focus only on joint information and bone information, but physical information such as the transformation angles of the human

DOI: 10.24507/icicelb.15.02.171

skeleton in the spatial domain and the motion trajectories in the temporal domain also contain rich features.

To address the above problem, we propose an attentional multi-scale unified spatiotemporal graph convolutional network (AMU-GCN). Firstly, inspired by Liu et al. [11], a multi-scale unified spatiotemporal graph convolutional network is proposed to solve the problem that spatiotemporal separated graph convolution cannot well establish multi-order cross-spatiotemporal connections of human joints. The introduction of an external attention mechanism [12] combined with multi-scale unified spatiotemporal graph convolution can not only effectively extract detailed information about the human skeleton in the action sequence, but also take account of the potential relationship between various actions. The 3D joint features are expanded into five high-level features, and each type of feature is input into the network to form a six-stream network to improve the performance of the network. The models are evaluated on two datasets, NTU RGB+D 60 and NTU RGB+D 120. Compared with previous mainstream methods, the accuracy of the AMU-GCN achieves more advanced results on both skeleton datasets.

The main contributions of this paper are as follows. 1) The multi-scale unified spatiotemporal graph convolution treats the skeleton sequence as a spatiotemporally crossable whole. Edge connections in the spatiotemporal skeleton are constructed by a sliding time window mechanism, to realize the cross-spatiotemporal graph convolution. Dividing the adjacency matrix into adjacency sub-matrices at each scale according to the shortest spatial distance ensures the validity of long-distance features. 2) The external attention mechanism extracts potential connections between different samples, balancing global and local features of the skeleton sequence. 3) A multi-input branch architecture is proposed to expand the original 3D joint features into five high-level features. The fusion of multi-stream information can help the network effectively integrate information data with different high-quality features and provide more accurate output.

The organization of this study is as follows. Section 2 reviews previous deep learning-based methods. Section 3 describes our proposed model in detail. Section 4 provides the experiment results and analysis. Section 5 summarizes the proposed model and indicates the future research direction.

2. **Related Work.**

Deep learning methods have achieved excellent results in human action recognition tasks and can be classified into three categories: recurrent neural network (RNNs) based, convolutional neural networks (CNNs) based, and graph convolutional neural networks (GCNs) based.

The RNNs-based approach feeds continuous skeleton data into the network as time series, focusing more on the extraction of temporal relationships. Liu et al. [13] combine the trust gate mechanism with long short-term memory (LSTM) and use multichannel fusion techniques to integrate features from each perspective of the action sequence. Liu et al. [14] use global context and attention mechanisms to capture key features of skeleton sequences. However, these methods focus more on temporal connections and fail to fully consider the extraction of spatial connections.

The CNNs-based approach processes the human skeleton as a spatiotemporal feature map to extract information. Kim and Reiter [15] stack multiple temporal convolutional layers to extract feature information in the skeleton. Liu et al. [16] map 3D skeleton data to 2D space, effectively eliminating viewing angle differences. However, these methods treat the skeleton information as a pseudo-image and ignore the 3D connections among the skeleton joints.

The GCNs-based approach [1-11,17-19] describes the connection status between nodes by constructing the adjacency matrix and updates the feature information of nodes by aggregation operations. Yan et al. [2] express the human skeleton sequence as a spatial graph and a temporal graph, respectively, and use the graph convolutional network to

model the human skeleton for the first time, achieving more advanced performance. Li et al. [3] represent skeleton sequences as action structure graphs, using graph convolutions to capture spatiotemporal features. At the same time, an attention mechanism is also introduced to weigh the nodes and connected edges. Shi et al. [4] divide the input features into two categories, joint features, and bone features, to capture the spatiotemporal relationship of the human body separately. Shi et al. [5] represent the human topological graph as a directed graph while introducing attention mechanisms and residual connections. Zhang et al. [6] map the skeleton information to a higher-level space and embed two kinds of semantic information, joint type, and frame index, to construct a lightweight model to effectively capture the differences between human actions. Song et al. [7] use the early fused multi-branch architecture and bottleneck residuals to realize the light weight of the model. Such methods are able to handle irregular topological data, taking full account of the connections among skeleton nodes.

3. Methods.

3.1. Unified spatiotemporal graph convolution.

The disadvantage of the spatiotemporal separation graph convolution is that it cannot directly extract the connections across space-time. As shown in Figure 1(a), joint 1 is very close to joint 2. However, these two joints do not belong to the same spatial dimension nor the same temporal dimension, so a spatial and temporal graph convolution is required to establish the connection. However, as the number of aggregation layers increases, more interference information is generated, weakening the connection between these two joints.

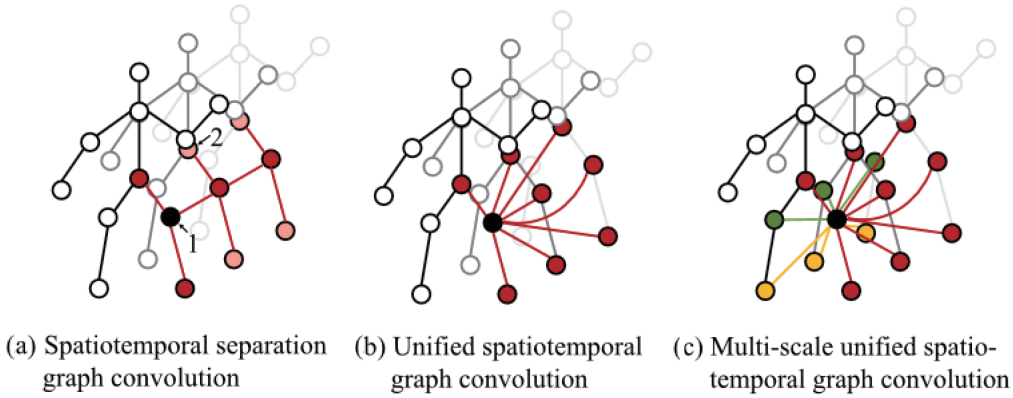

FIGURE 1. Comparison of three graph convolution methods

Unified spatiotemporal graph convolution enables cross-spatiotemporal connectivity, as shown in Figure 1(b). First, a sliding time window is given in the time dimension, and the window size is set to s. Each sliding step of the time window yields a spatiotemporal subgraph $G_s = (V_s, E_s)$ based on the skeleton nodes and neighboring edges. V_s represents the union of each frame joint point set in the s frame, and the initialization form of E_s is the adjacency matrix $\tilde{\mathbf{A}}_s$, and the specific expression is as follows:

$$\tilde{\mathbf{A}}_s = \begin{matrix} \tilde{\mathbf{A}} & \cdots & \tilde{\mathbf{A}} \\ \vdots & \ddots & \vdots \\ \tilde{\mathbf{A}} & \cdots & \tilde{\mathbf{A}} \end{matrix} \tag{1}$$

where $\tilde{\mathbf{A}}$ is the adjacency matrix of skeleton joint points in a single frame, and the size is $M \times M$. $\tilde{\mathbf{A}}_s$ is the tiling of $\tilde{\mathbf{A}}$ in each frame, with a total of s frames, so the size is $sM \times sM$. $\left[\tilde{\mathbf{A}}_s\right]_{ij} = \tilde{\mathbf{A}}$ means that each node in V_i can be connected to the adjacent nodes of the same node in frame j. All joint points not only have spatial connections within a

single frame, but also have direct connections with the adjacent one-hop neighbors of the same node in the s frame. From this, the sliding feature $\mathbf{X}_s \in \mathbb{R}^{T\times sM\times C}$ can be obtained. Then use $\tilde{\mathbf{A}}_s$ as the adjacency matrix to perform cross-spatiotemporal graph convolution on $\mathbf{X}_s$. The details are as follows:

$$(\mathbf{X}_{\text{out}})_s = \sigma\left[\tilde{\mathbf{D}}_s^{-\frac{1}{2}}\tilde{\mathbf{A}}_s\tilde{\mathbf{D}}_s^{-\frac{1}{2}}(\mathbf{X}_{\text{in}})_s\mathbf{W}_{\text{in}}\right] \tag{2}$$

where $\tilde{\mathbf{D}}_s$ is the degree matrix of $\tilde{\mathbf{A}}_s$, $\mathbf{W}$ is the weight parameter, and σ is the ReLU activation function.

3.2. **Multi-scale graph convolution.** To establish connections between neighboring nodes with different step sizes, multi-scale graph convolution is designed. The multi-scale adjacency matrix is defined as follows:

$$\left(\tilde{\mathbf{A}}_p\right)_{i,j} = \begin{cases} 1, & \text{if } d(\mathbf{v}_i,\mathbf{v}_j) = p \\ 1, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where $d(\mathbf{v}_i,\mathbf{v}_j)$ represents the shortest hop distance between $\mathbf{v}_i$ and $\mathbf{v}_j$. Multi-scale adjacency matrix $\tilde{\mathbf{A}}_p$ obtained by setting different p values, the weights of nodes of different scales are not affected by nodes of other scales. The representation of multi-scale graph convolution is as follows:

$$\mathbf{X}_{\text{out}} = \sigma\left[\sum_{p=0}^{P}\tilde{\mathbf{D}}_p^{-\frac{1}{2}}\tilde{\mathbf{A}}_p\tilde{\mathbf{D}}_p^{-\frac{1}{2}}\mathbf{X}_{\text{in}}(\mathbf{W}_{\text{in}})_p\right] \tag{4}$$

The convolution of graphs at different scales is aggregated in an additive manner in order to make the information of nodes with different p values valid for a long time. Extending multi-scale graph convolution to unified spatiotemporal graph convolution can result in multi-scale unified spatiotemporal graph convolution, as shown in Figure 1(c). The calculation method of multi-scale unified spatiotemporal graph convolutional network is shown as follows:

$$(\mathbf{X}_{\text{out}})_s = \sigma\left[\sum_{p=0}^{P}\tilde{\mathbf{D}}_{s,p}^{-\frac{1}{2}}\tilde{\mathbf{A}}_{s,p}\tilde{\mathbf{D}}_{s,p}^{-\frac{1}{2}}[\mathbf{X}_{\text{in}}]_s(\mathbf{W}_{\text{in}})_p\right] \tag{5}$$

3.3. **External attention.** The external attention mechanism shares the parameters of each sample through two small external shared memories and explores the potential connections between different samples to improve the generalization ability of the entire model. The input skeleton sequence is characterized as $\mathbf{X} \in \mathbb{R}^{C\times T\times V}$, where C is the channel dimension, T is the number of frames of a single complete action sequence, and V is the number of nodes in the skeleton topology map. The two independent input memories are $\mathbf{M}_{\text{key}} \in \mathbb{R}^{Q\times C}$ and $\mathbf{M}_{\text{value}} \in \mathbb{R}^{Q\times C}$, where Q is the number of elements. First, transform the input features to obtain $\mathbf{X}_{\text{query}} \in \mathbb{R}^{N\times C}$, where $N = T \times V$. The attention map $\tilde{\mathbf{H}}$ is then generated by computing the correlation between $\mathbf{X}_{\text{query}}$ and the memory $\mathbf{M}_{\text{key}}$. Finally, after the double normalization operation, the attention map $\mathbf{H}$ is multiplied with another memory $\mathbf{M}_{\text{value}}$ to obtain the new feature map. The specific process is shown in Figure 2. The calculation is as follows:

$$\tilde{\mathbf{H}} = \mathbf{X}_{\text{query}}\mathbf{M}_{\text{key}}^{\text{T}} \tag{6}$$

$$\mathbf{H} = \text{doubleNorm}\left(\tilde{\mathbf{H}}\right) \tag{7}$$

$$\mathbf{X}_{\text{out}} = \mathbf{H}\mathbf{M}_{\text{value}} \tag{8}$$

$\mathbf{M}_{\text{key}}$ and $\mathbf{M}_{\text{value}}$ are learnable. Since there is no prior, both can be implemented using linear layers and optimized by end-to-end backpropagation. The doubleNorm is a double

normalization operation. Usually, only softmax is used in the algorithm to normalize N. However, note that the attention map is obtained by channel dimension transformation, which is also sensitive to channel information, so the L1 normalization operation is performed on the channel dimension.

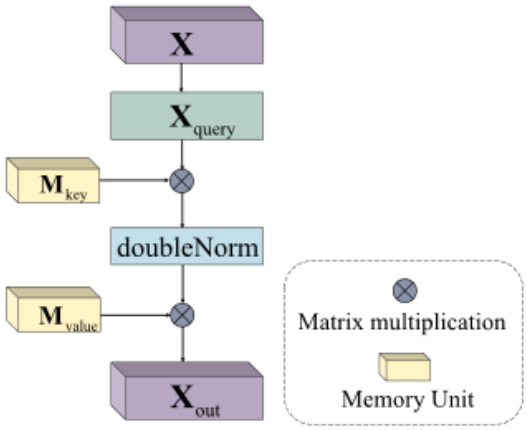

FIGURE 2. External attention mechanism process

The multi-scale unified spatialtemporal graph convolution (MUGC) module and the external attention (EA) module have no order in the training process. As shown in Figure 3, each layer of the network inputs features into the two modules at the same time. Finally, the features learned by the two modules are added together and then input into the next layer network.

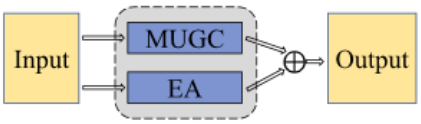

FIGURE 3. Attentional multi-scale unified spatiotemporal graph convolution

3.4. **Multi-stream architecture.** Extending the original features into more new high-quality features can effectively improve the performance of the model. We divide the features into 6 types of inputs.

1) Joint information. The 3D coordinate information of the skeleton nodes, denoted as $\mathbf{v} = (x, y, z)$.

2) Relative joint information. Find a central joint point in the skeleton and represent the joints of the human body as a coordinate system with variations. Given that the central node at frame t is $\mathbf{v}_{c,t} = (x_{c,t}, y_{c,t}, z_{c,t})$ and the another node is $\mathbf{v}_{i,t} = (x_{i,t}, y_{i,t}, z_{i,t})$, the relative joint information can be expressed as $\mathbf{r}_{i,t} = (x_{i,t} - x_{c,t}, y_{i,t} - y_{c,t}, z_{i,t} - z_{c,t})$.

3) Bone information. The difference between adjacent joint points indicates the length and direction of the bone. Given that the source node at frame t is $\mathbf{v}_{i,t} = (x_{i,t}, y_{i,t}, z_{i,t})$ and the target node is $\mathbf{v}_{j,t} = (x_{j,t}, y_{j,t}, z_{j,t})$, the bone information can be expressed as $\mathbf{e}_{\mathbf{v}_{i,t},\mathbf{v}_{j,t}} = (x_{j,t} - x_{i,t}, y_{j,t} - y_{i,t}, z_{j,t} - z_{i,t})$.

4) Bone angle information. The angle of each bone can more accurately extract subtle changes in human motion. Given that the bone information at frame t is $\mathbf{e}_{i,t} = (x_{i,t}, y_{i,t}, z_{i,t})$, the bone angle information can be expressed as $\mathbf{a}_{i,t} = \arccos\left(\mathbf{e}_{i,t} \Big/ \sqrt{x_{i,t}^2 + y_{i,t}^2 + z_{i,t}^2}\right)$.

5) Fast time information. The high frame rate motion information is the displacement of the same joint in consecutive adjacent frames, which helps to capture some local and fast motion details. Given that the coordinate information of the joints at frame t is $\mathbf{v}_{i,t} = (x_{i,t}, y_{i,t}, z_{i,t})$, and the coordinate information of the same joints at frame $t + 1$ is

$\mathbf{v}_{i,t+1} = (x_{i,t+1}, y_{i,t+1}, z_{i,t+1})$, the fast time information can be expressed as $\mathbf{mf}_{i,t,t+1} = (x_{i,t+1} - x_{i,t}, y_{i,t+1} - y_{i,t}, z_{i,t+1} - z_{i,t})$.

6) Slow time information. The low frame rate motion information is the difference between two adjacent frames of the same joint, which helps to capture the action with a strong overall correlation. Given that the coordinate information of the joints at frame t is $\mathbf{v}_{i,t} = (x_{i,t}, y_{i,t}, z_{i,t})$, and the coordinate information of the same joints at frame $t+2$ is $\mathbf{v}_{i,t+2} = (x_{i,t+2}, y_{i,t+2}, z_{i,t+2})$, the slow time information can be expressed as $\mathbf{ms}_{i,t,t+2} = (x_{i,t+2} - x_{i,t}, y_{i,t+2} - y_{i,t}, z_{i,t+2} - z_{i,t})$.

3.5. **Overall framework.** The overall framework of the AMU-GCN is shown in Figure 4. The network is divided into six streams of input, and after three layers of attentional multi-scale unified spatiotemporal graph convolution (AMU-GC) module, the six streams of recognition results are fused according to the corresponding weight ratio to output the final action classification. The initial number of channels is 3, and the number of channels output by each layer of AMU-GC module is 96, 192 and 384, respectively.

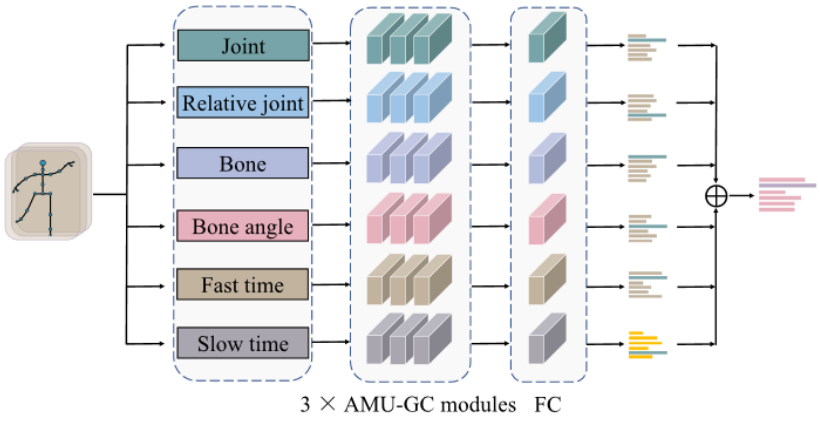

FIGURE 4. Overall framework of AMU-GCN

4. Main Results.

4.1. **Ablation experiments.** The model is compared and evaluated on two 3D human skeleton datasets, NTU RGB+D 60 and NTU RGB+D 120. We set the unified spatiotemporal graph convolution as a dual path, and the sliding window s is initialized to two values, one of which is set to 3 to capture short-distance spatiotemporal features, and the other is set to 5 to capture long-distance spatiotemporal features. The value of the multi-scale graph convolution parameter P is set to 5 and Q is set to 64 in the external attention. As shown in Table 1, it can be seen that the multi-scale unified spatiotemporal graph convolution is improved compared with the baseline spatial temporal graph convolutional network (ST-GCN) [2]. Combining the external attention mechanism with multi-scale unified spatio-temporal graph convolution, the expressive ability of the model is strengthened, and the experiment accuracy is improved to a certain extent.

As shown in Table 2, 1s indicates that the input feature is joint information, and 6s indicates that the input features are joint information, relative joint information, bone

TABLE 1. Model accuracy under different configurations on NTU RGB+D 60 dataset

Methods	X-sub (%)	X-view (%)
Baseline ST-GCN [2]	84.2	89.7
AMU-GCN w/o EA	84.6	92.2
AMU-GCN	**86.5**	**93.0**

information, bone angle information, fast time information, and slow time information. The experiment results show that the accuracy of the model improves with each additional stream of feature information.

TABLE 2. Multi-stream architecture comparison experiments on NTU RGB +D 60 and NTU RGB+D 120 datasets

Methods	X-sub (%)	X-view (%)	X-sub120 (%)	X-set120 (%)
1s AMU-GCN	86.5	93.0	77.9	82.2
2s AMU-GCN	89.5	94.7	84.5	87.0
3s AMU-GCN	90.1	95.4	84.7	87.8
4s AMU-GCN	90.4	95.5	84.9	88.1
5s AMU-GCN	90.7	95.6	85.1	88.2
6s AMU-GCN	**90.8**	**95.7**	**85.3**	**88.6**

4.2. **Comparison with previous models.** The model is evaluated on two public datasets, NTU RGB+D 60 and NTU RGB+D 120, and compared with other action recognition methods, as detailed in Table 3. Compared with RNNs-based [14], CNNs-based [17] and GCNs-based [2-6,9] methods, AMU-GCN shows a significant improvement in accuracy. Experiment results show that the AMU-GCN has superior performance.

TABLE 3. Comparisons with different models on NTU RGB+D 60 and NTU RGB+D 120 datasets

Methods	X-sub (%)	X-view (%)	X-sub120 (%)	X-set120 (%)
ST-LSTM [14]	69.2	77.7	55.7	57.9
Synthesized CNN [17]	80.0	87.2	–	–
ST-GCN [2]	81.5	88.3	70.7	73.2
2s AS-GCN [3]	86.8	94.2	77.9	78.5
2s AGCN [4]	88.5	95.1	84.2	85.5
SGN [6]	89.0	94.5	79.2	81.5
MSSF-GCN [9]	89.5	**96.2**	84.4	86.1
4s DGNN [5]	89.9	96.1	–	–
6s AMU-GCN (ours)	**90.8**	95.7	**85.3**	**88.6**

5. **Conclusions.** In this paper, we propose an attentional multi-scale unified spatiotemporal graph convolutional network. First, the multi-scale unified spatiotemporal graph convolution not only realizes the barrier-free spatiotemporal graph convolution but also ensures the validity of different scale features. Meanwhile, the external attention mechanism can learn more representative features from the whole sample through the external memory. In addition, multi-stream architecture integrates various high-quality features to build a reliable model and provide more accurate action recognition information. Compared with other mainstream methods, the model shows excellent performance on public datasets NTU RGB+D 60 and NTU RGB+D 120. Future research directions will focus on how to achieve lightweight models.

Acknowledgment. This study was supported in part by the Natural Science Foundation of Xinjiang Uygur Autonomous Region (No. 2022D01A59), the National Natural Science Foundation of China (No. U20A20167), and the Innovation Capability Improvement Plan Project of Hebei Province (No. 22567637H).

REFERENCES

[1] M. Fajar, Y. Udjaja, David, A. Chowanda, B. Juarto and Yulan, A comparative investigation of usability issues toward virtual reality implementation in a state-owned shipping service enterprise, *ICIC Express Letters, Part B: Applications*, vol.13, no.5, pp.545-552, 2022.

[2] S. Yan, Y. Xiong and D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, *Processdings of the AAAI Conference on Artificial Intelligence*, vol.32, no.1, 2018.

[3] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang and Q. Tian, Actional-structural graph convolutional networks for skeleton-based action recognition, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, pp.3595-3603, 2019.

[4] L. Shi, Y. Zhang, J. Cheng and H. Lu, Two-stream adaptive graph convolutional networks for skeleton-based action recognition, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, pp.12026-12035, 2019.

[5] L. Shi, Y. Zhang, J. Cheng and H. Liu, Skeleton-based action recognition with directed graph neural networks, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, pp.7912-7921, 2019.

[6] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue and N. Zheng, Semantics-guided neural networks for efficient skeleton-based human action recognition, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, pp.1112-1121, 2020.

[7] Y. Song, Z. Zhang, C. Shan and L. Wang, Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition, *Proc. of the 28th ACM International Conference on Multi-Media*, Seattle, pp.1625-1633, 2020.

[8] X. Zhang, C. Xu and D. Tao, Context aware graph convolution for skeleton-based action recognition, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, pp.14333-14342, 2020.

[9] N. Sun, L. Leng, J. Liu and G. Han, Multi-stream slowfast graph convolutional networks for skeleton-based action recognition, *Image and Vision Computing*, vol.109, 104141, 2021.

[10] Y. Zang, D. Yang, T. Liu, H. Li, S. Zhao and Q. Liu, SparseShift-GCN: High precision skeleton-based action recognition, *Pattern Recognition Letters*, vol.153, pp.136-143, 2022.

[11] Z. Liu, H. Zhang, Z. Chen, Z. Wang and W. Ouyang, Disentangling and unifying graph convolutions for skeleton-based action recognition, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, pp.143-152, 2020.

[12] M. Guo, Z. Liu, T. Mu and S. Hu, Beyond self-attention: External attention using two linear layers for visual tasks, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[13] J. Liu, A. Shahroudy, D. Xu and G. Wang, Spatio-temporal LSTM with trust gates for 3D human action recognition, *Proc. of the 14th European Conference on Computer Vision (ECCV2016)*, Amsterdam, The Netherlands, pp.816-833, 2016.

[14] J. Liu, G. Wang, P. Hu, L. Duan and A. C. Kot, Global context-aware attention LSTM networks for 3D action recognition, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp.3671-3680, 2017.

[15] T. S. Kim and A. Reiter, Interpretable 3D human action analysis with temporal convolutional networks, *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Hawaii, pp.1623-1631, 2017.

[16] M. Liu, H. Liu and C. Chen, Enhanced skeleton visualization for view invariant human action recognition, *Pattern Recognition*, vol.68, pp.346-362, 2017.

[17] W. Peng, X. Hong, H. Chen and G. Zhao, Learning graph convolutional network for skeleton-based human action recognition by neural searching, *Processdings of the AAAI Conference on Artificial Intelligence*, New York, vol.34, no.3, pp.2669-2676, 2020.

[18] Y. Jiang, X. Yang, J. Liu and J. Zhang, A lightweight hierarchical model with frame-level joints adaptive graph convolution for skeleton-based action recognition, *Security and Communication Networks*, vol.2021, pp.1-13, 2021.

[19] H. Chen, M. Li, L. Jing and Z. Cheng, Lightweight long and short-range spatial-temporal graph convolutional network for skeleton-based action recognition, *IEEE Access*, vol.9, pp.161374-161382, 2021.