Multi-scale Dilated Attention Graph Convolutional Network for Skeleton-Based Action Recognition

Yang Shu, Wanggen Li(✉), Doudou Li, Kun Gao, and Biao Jie

Anhui Normal University, Wuhu, China
{yangshu,xchen}@ahnu.edu.cn

Abstract. Due to the small size, anti-interference and strong robustness of skeletal data, research on human skeleton-based action recognition has become a mainstream. However, due to the incomplete utilization of semantic information and insufficient time modeling, most methods may not be able to fully explore the connections between non-adjacent joints in the spatial or temporal dimensions. Therefore, we propose a Multi-scale Dilated Attention Graph Convolutional Network for Skeleton-Based Action Recognition (MDKA-GCN) to solve the above problems. In the spatial configuration, we explicitly introduce the channel graph composed of high-level semantics (joint type and frame index) of joints into the network to enhance the representation ability of spatiotemporal features. MDKA-GCN uses joint-level, velocity-level and bone-level graphs to more deeply mine the hidden features of human skeletons. In the time configuration, two lightweight multi-scale strategies are proposed, which can be more robust to time changes. Extensive experiments on NTU-RGB+D 60 datasets and NTU-RGB+D 120 datasets show that MDKA-GCN has reached an advanced level, and surpasses the performance of most lightweight SOTA methods.

Keywords: Action Recognition · Multi-scale · Semantic Information · Dilated Attention · Lightweight

1 Introduction

In recent years, the task of action recognition has become one of the most attractive topics in the field of artificial intelligence, especially human action recognition (HAR) is widely used in various fields such as human or object interaction, video surveillance systems and healthcare systems [1], providing accurate judgment analysis and understanding of human actions for machinery and equipment in these fields, playing a crucial role in the development and progress of artificial intelligence.

Early on, research on human skeleton action recognition is mainly through deep neural network models to learn the correlation of human actions in time and space. In these models, the performance of human skeleton action recognition

Q. Liu et al. (Eds.): PRCV 2023, LNCS 14425, pp. 16–28, 2024.
https://doi.org/10.1007/978-981-99-8429-9_2

based on graph convolutional networks (GCN) [2] is better than that based on recurrent neural networks (RNN) [3] and convolutional neural networks (CNN) [4]. GCN methods can construct a spatiotemporal topology graph of 3D positions of human skeleton joint nodes by regarding human joint nodes as vertices of a graph, treating natural topological connections between adjacent joint nodes as spatial edges of a graph and considering temporal correlation between adjacent frames as temporal edges. Then input the processed human skeleton topology graph sequence into the network for learning to finally achieve action classification. The GCN-based method has been proven to be an effective solution for achieving the task of human action recognition.

To further improve the performance of the model, they [5–8] focus on introducing adaptive graph residual masks to capture the relationships between different joints, that is, to extract more hidden information from the original human skeleton dataset, such as bone and velocity. In order to enhance the feature representation of every actions, they train this information through multiple network streams and fuse all the trained features together to obtain the score of each action and achieve the task of action classification. However, more information will cause information redundancy and model size doubling sacrificing model storage space and computational efficiency, which is extremely disadvantageous for model promotion in practical applications.

In response, SGN [9] achieves superior performance with a smaller model, however, it also has problems such as insufficient data mining and incomplete semantic utilization. Guided by literature [10,11], we consider combining channel attention with dilated convolution attention to enhance feature connections between frame dimensions and channels in the model. The main contributions of this paper can be summarized as follows:

- This paper introduces multiple hidden information of human skeletons after data preprocessing and effectively fuses them in the early stage of the model, enhancing feature representation of each information and obtaining a richer topology graph.
- In order to make full use of two semantic relationships, we integrate two semantic information into graph convolution modules by adjusting graph convolution layers and channel width effectively, solving defects in spatial-temporal separation processing.
- In the time module, we design a time multi-scale dilated convolution kernel attention (T-MDKA), to obtain a large receptive field by replacing large kernel convolution with dilated convolution, thereby simulating remote dependencies. In addition, we construct two branch time convolution blocks to more robustly learn the temporal features of actions.

2 Related Works

2.1 Attention Mechanism

The attention mechanism can be seen as simulating the degree of attention that people pay to a certain part when processing information by adjusting the size

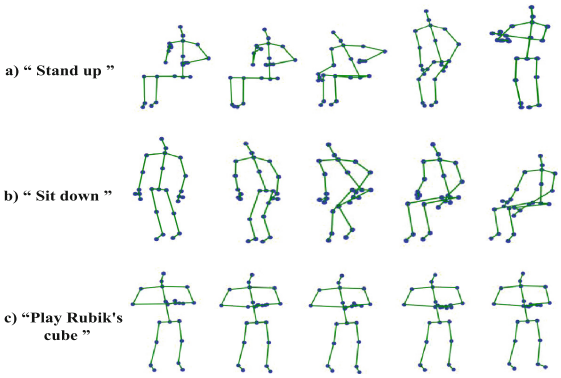

Fig. 1. Skeleton diagrams of 5 frames from three action sequences.

of the weights. It is currently widely used in various fields. [10–12]. SENet [10] proposes a squeeze-and-excitation block to learn global channel information, which enables the model to focus on more useful feature information. VAN [12] improves channel adaptability by using large kernels. To maximize the role of large convolution kernels, MAN [11] adopts the structure of transformer and introduces GSAU to replace MLP structure to obtain multi-scale remote modeling dependencies not only improving model representation ability, but also reducing model parameters and computational complexity.

2.2 Lightweight Models

In the image field [13] and object detection field [14], methods using depth-separable convolution and grouped convolution are proposed respectively to replace traditional convolution greatly reducing model parameters. Zhang et al. [9] based on graph convolutional neural networks introduce high-order semantic information to enhance feature expression ability achieving low parameters while maintaining high recognition accuracy. Cheng et al. [15] construct a lightweight network framework using dynamic displacement graph convolution instead of traditional convolution, In order to further simplify, they [16] use edge RELU distillation technology, which also improves model recognition performance. In addition, Song et al. [17] embed separable convolution layers into early multi-information fusion module. It makes the model's parameter size extremely small, making the model more lightweight.

3 Method

In this section, we will detail the composition of our proposed MDKA-GCN. Figure 2(a) is our overall model framework.

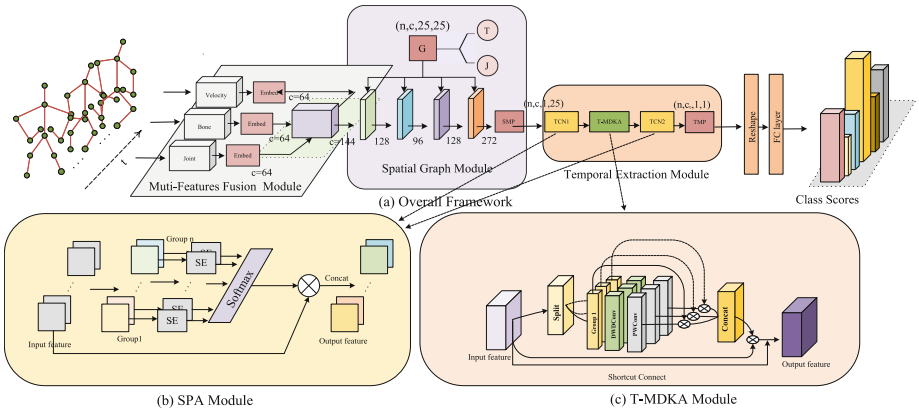

Fig. 2. (a) the overall framework of this paper's model includes three parts; (b) the pyramid partition attention module applied to TCN1 and TCN2; (c) the dilated convolution attention module proposed in this paper.

3.1 Multi-Branch Fusion Module

Earlier research [17] has shown that more skeleton topology graphs play a key role in model performance. In this work, we mine three types of input features from skeleton data: 1) joint stream, 2) velocity stream, 3) bone stream.

Specifically, this paper represents the skeleton sequence as a set of joint sets $\mathbf{S}_{t,k} = \{x_{t,k} | t = 1, 2, ..., T; k = 1, 2, ..., J\}$, where T represents the total number of time frames and J represents the total number of human joints.

Joint $\mathbf{S}_{t,k} \in \mathbb{R}^{C \times T \times V}$ is the original 3D coordinate provided by the datasets where channel C is equal to 3. Therefore, through formula $\mathrm{v}_{t,k} = \mathrm{x}_{t,k} - \mathrm{x}_{t,m}$, we can obtain the relative position of joints where $\mathrm{x}_{t,m}$ represents the position of the human skeleton's center of gravity. Considering that many subtle actions are concentrated on the hands such as "play Rubik's cube" in Fig. 1(c), we determine three central joints as the upper and lower spine and palm wrist joints of the action sequence.

Similarly joint velocity can also be easily defined by joint position that is the position change between adjacent frames represented as $\mathrm{v}_{t,k} = \mathrm{x}_{t,k} - \mathrm{x}_{t-1,k}$.

Like the definition of relative position, bone information can also be defined by the position difference between two adjacent joints on a skeleton represented as $\mathrm{b}_{t,k} = \mathrm{x}_{t,k} - \mathrm{x}_{t,i}$, where joint $\mathrm{x}_{t,k}$ represents a position away from the human body's center of gravity and joint $\mathrm{x}_{t,i}$ represents a position close to the human body's center of gravity adjacent to joint $\mathrm{x}_{t,k}$. Similar to formula of joint velocity, we can easily get bone velocity. Since acceleration information is crucial for capturing some small actions, we can obtains bone acceleration information from bone velocity information as input information. Finally, these input information are encoded through two fully connected layers (FC),

$$\mathrm{P}_{t,k} = \sigma(\mathrm{FC}(\sigma(\mathrm{FC}(\mathrm{x}_{t,k})))) \tag{1}$$

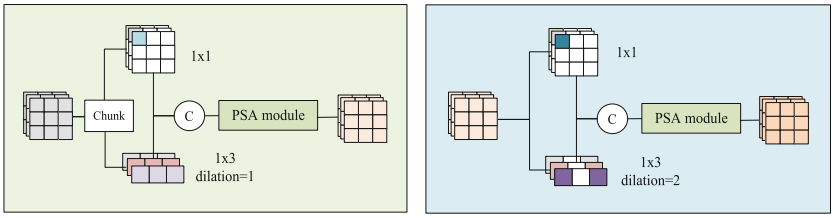

Fig. 3. (a) and (b) are two divergent convolution modules that integrate the Pyramid Split Attention (PSA) module.

where $\mathrm{P}_{t,k}$ represents the joint information encoded by the fully connected layer; σ in this paper represents activation function RELU.

In the feature fusion module, we obtain different channel attention weights through two layers of convolution operations, further enhancing feature representation of each information stream. Then fuse three information streams and input them into graph convolution modules,

$$\hat{\mathrm{P}}_{t,k} = Conv(\sigma(bn(\mathrm{p}_{t,k}))) \tag{2}$$

$$f_{\mathrm{in}} = Cat[\hat{\mathrm{P}}_{t,k}, \hat{\mathrm{V}}_{t,k}, \hat{\mathrm{B}}_{t,k}] \tag{3}$$

where $\hat{\mathrm{V}}_{t,k}$ and $\hat{\mathrm{B}}_{t,k}$ represents respectively the velocity information and bone information encoded by the fully connected layer; bn and Cat represent normalization function and concatenation operation respectively. $Conv$ represents pointwise convolution, which reduces the channel dimension to avoid generating a large number of parameters due to high dimensions after concatenation.

3.2 Semantic Information

In this work, we use a one-hot vector J'_k to represent the kth skeleton joint. Similarly use a one-hot vector T'_i to represent the ith frame index. This paper concatenates two semantic information in low dimensions, then convolves them through multi-layer perceptron (MLP). Finally, input them into each layer graph convolution,

$$\mathrm{G}_0 = Cat[\mathrm{J}'_k, \mathrm{T}'_i] \tag{4}$$

$$\mathrm{G}_j = \mathrm{MLP}(\mathrm{G}_{j-1}) \tag{5}$$

where J'_k and T'_i respectively refer to joint type and frame index semantic information, and $j = 1, 2, 3, 4$. The purpose of MLP is to increase channel dimensions to match input dimensions of each layer graph convolution.

3.3 Graph Convolution Module

The operation in this paper's graph convolution module is different from previous work [9], which often extracts spatial features through adjacency matrices

composed of natural human joint nodes or deformations thereof. This paper considers the importance of two semantic information, inputs channel graphs fused with semantics into graph convolution modules,

$$f_{\text{out}} = \sigma(bn(Conv(f_{\text{in}} \otimes \mathrm{G}_j) + Conv(f_{\text{in}}))) \tag{6}$$

where f_{in} and f_{out} are respectively input and output of graph convolution, $\otimes$ represents matrix multiplication, the size of the convolutional kernel of $Conv$ is 1×1 with different training weights.

3.4 Time Convolution Module

Considering that time processing method of SGN only uses a fixed convolution operation which is not enough to distinguish some similar actions, such as Fig. 1(a) and Fig. 1(b). Therefore, we propose a time convolution module composed of two branch convolution blocks in Fig. 3 and multi-scale dilated attention in Fig. 2(c).

Branch Convolution Block. In order to obtain receptive fields at different scales while controlling model parameter volume and computational volume. Inspired by [18], We introduce the PSA module and design two types of pyramid split convolution modules in Fig. 2(b) for extracting multi-scale temporal features of action sequences,

$$F_i = Conv_i(\sigma(bn(chunk(\mathrm{X}_{\text{in}}))) \tag{7}$$

$$F = F_0 \oplus F_1 \tag{8}$$

where *chunk* is the split operator, which divides the input information into two equal parts on the channel dimension. And $Conv_i$ means the convolution operation with kernel sizes of 1×1 and 1×3. $\oplus$ is the concat operator. Then, the SEweight module is used to obtain the attention weight from the input feature map with different scales,

$$Z_i = \text{SEweight}(F_i) \tag{9}$$

$$Z = \text{Softmax}(Z_0 \oplus Z_1) \tag{10}$$

$$Out = F \odot Z \tag{11}$$

where the $Softmax$ is used to obtain the re-calibrated weight, $\odot$ represents the dot product operation. Different from $TCN1$, $TCN2$ uses grouped convolutions and dilated convolutions with dilation rate 2 with kernel size 1×3.

Time Multi-scale Dilated Attention. Inspired by some large kernel works [8,11] without adding too much computational burden obtaining advantages of attention mechanisms in long-term modeling, we propose a time multi-scale

dilated kernel attention and replace large kernel convolutions with dilated convolutions focusing on extracting temporal dependencies between action sequences. It is shown in Fig. 2(c), the formula is as follows:

$$x_i = Split(x) \tag{12}$$

$$DKA(x_i) = PWConv(Conv_{DWD}(x_i)) \tag{13}$$

$$MDKA(x) = Cat(DKA(x_i) \odot x_i) \tag{14}$$

where $Split$ is the split operation, $i = 1, 2, 3, 4$; $Conv_{DWD}$ is a dilated separable convolution with a kernel size of 1×3, and the dilation rate can be $2, 3, 4$, $PWConv$ represents a normal pointwise convolution.

4 Experiment

4.1 Dataset

NTU-RGB+D 60 Dataset [19]: One of the current mainstream skeleton-based action recognition datasets, containing 56880 skeleton sequences of 60 action categories captured simultaneously from 40 different subjects and 3 Microsoft Kinect V2 depth cameras. Each skeleton sequence contains three-dimensional spatial coordinates of 25 joints. The dataset provides two evaluation benchmarks: Cross-Subject (C-Sub) and Cross-View (C-View). C-Sub is completed by 40 subjects with half of the subjects used for training and the rest for testing. C-View selects samples captured by cameras 2 and 3 for training and the rest for testing.

NTU-RGB+D 120 Dataset [20]: This dataset is an expansion of the NTU RGB+D 60 dataset in terms of action categories and number of actors, containing 114480 action videos participated by 106 actors, with a total of 120 action categories including 82 daily life actions 12 medical conditions and 26 actions under two-person interaction. The dataset has two evaluation benchmarks: Cross-Subject (C-Sub120) and Cross-Setup (C-Set120). C-Sub120 divides this dataset into training set (63026 videos) and validation set(50919 videos) according to different actors in the video. C-Set120 divides the dataset according to the parity of video numbers. 54468 even-numbered videos are used as training sets, and 59477 odd-numbered videos are used as test sets.

4.2 Experimental Details

Similar to [9], the difference is that in order to facilitate the operation of channel graphs composed of two semantic information in the graph convolution module, we adjust the time dimension to 25. This paper sets the number of epochs in the model to 120 sets the batch size for each epoch to 64 sets, the initial learning rate to 0.001 and continues to decrease during iteration. When the number of iterations is 80 and 100 the learning rate drops tenfold. At the same time, this paper also uses Adam to optimize the model where weight decay is 0.0001.

4.3 Ablation Experiment

In this part, we mainly discuss the contributions of different components in this paper's model, which includes multi-branch fusion module, high-order semantic information time convolution module and necessity of attention module.

Table 1. Comparison of the accuracy of different input branches.

Input	Param(M)	C-Sub/%	C-View/%
Joint	0.60	88.4	94.4
Bone	0.60	89.8	94.8
Velocity	0.60	84.8	91.0
Joint and Bone	0.65	90.3	95.6
Joint and Velocity	0.65	90.4	95.8
Bone and Velocity	0.65	90.5	95.7
Baseline	0.69	**91.2**	**96.2**

Table 1 experimental results two conclusions are verified: first using information fusion of three input branches has obviously highest recognition accuracy on C-Sub and C-View; second this paper's information fusion method only increases about 1/7 of parameter volume while model performance has been significantly improved.

Table 2. Verification of the accuracy of two types of semantic information.

Input	Param(M)	C-Sub/%	C-View/%
w/o J	0.69	90.2	95.5
w/o T	0.69	90.5	95.3
w/o J and T	0.69	89.9	94.8
Baseline	0.69	**91.2**	**96.2**

Table 2 experimental results show the channel graph composed of two semantic information plays an important role in graph convolution operation. It is worth noting that when there is no channel graph the original graph convolution becomes ordinary pointwise convolution and the model's ability to aggregate different joint features will weaken resulting in a decline in model performance.

Table 3 verifies the effectiveness of the two branch convolution blocks and multi-scale dilated attention module. From the table it can be seen that the attention module proposed in this paper effectively improves model performance while adding a very small amount of parameters; and our reasonable combination of two branch convolution modules makes the model more balanced in terms of performance and parameter volume.

Table 3. Verification of the effectiveness of two divergent convolution modules and attention mechanisms.

Input	Param(M)	C-Sub/%	C-View/%
w/o Attention	0.70	90.7	96.0
w/o TCN1	0.62	90.9	95.8
w/o TCN2	0.55	90.8	95.4
TCN1 + TCN1	0.62	90.8	95.8
TCN2 + TCN2	0.77	90.8	96.0
Baseline	0.69	**91.2**	**96.2**

4.4 Comparison with State-of-the-Art

From Table 4, it can be seen that the best performance of our single-stream network MDKA-GCN (1s) on the two benchmarks is 91.2% and 96.2% respectively, while the establishment of multi-stream network enables the model to achieve better performance especially MDKA-GCN (4s) recognition accuracy on the two benchmarks respectively reach 92.1% and 96.8%, which is better than other SOTA models.

Table 4. Comparison of accuracy (%) with some recent SOTA methods.

Methods	Year	C-Sub/%	C-View/%	C-Sub120/%	C-Set120/%
ST-GCN [2]	2018	81.5	88.3	70.7	73.2
2s-AGCN [5]	2019	88.5	95.1	82.5	84.2
4s-Shift-GCN [15]	2020	90.7	96.5	85.9	87.6
SGN [9]	2020	89.0	94.5	79.2	81.5
MS-G3D [21]	2020	91.5	96.2	86.9	88.4
FGCN [22]	2021	90.2	96.3	85.4	87.4
CDGC [23]	2021	90.9	96.5	86.3	87.8
4s-Shift-GCN++ [16]	2021	90.5	96.3	85.6	87.2
Ta-CNN [4]	2022	90.7	95.1	85.7	87.3
4s-AGE-Ens [6]	2022	91.6	96.3	88.2	89.2
2s-ST-GCN++ [24]	2022	91.4	96.7	87.0	89.1
EffiencientB4 [17]	2022	**92.1**	96.1	**88.7**	88.9
SMotif-GCN [7]	2022	91.7	96.7	88.4	88.9
LKA-GCN [8]	2023	90.7	96.1	86.3	87.8
MDKA-GCN(1s)	–	91.2	96.2	86.8	88.3
MDKA-GCN(2s)	–	91.6	96.6	87.5	89.1
MDKA-GCN(4s)	–	**92.1**	**96.8**	87.9	**89.4**

Due to the randomness in the network during training, such as sample shuffling operations and frame extraction randomness, these random operations can cause incomplete feature learning in model training, resulting in unstable training results. To avoid this randomness and enhance the robustness of the network, we design a multi-stream network structure where each single-stream subnetwork structure is completely consistent. We fuse the output results of multiple single-stream sub-networks by adding them together and use them as the final output result of the multi-stream network.

In addition, Table 4 shows that on the C-Sub120 and C-Set120 compared with SGN [9], our single-stream method increases the accuracy by 7.6 percentage points and 6.8 percentage points respectively. In multi-stream networks (2s-AGCN [5], 4s-AGE-Ens [6], SMotif-GCN [7]), our multi-stream method MDKA-GCN(4s), although lower in accuracy on C-Sub120 than 4s-AGE-Ens and SMotif-GCN, reach the highest accuracy on C-Set120 benchmark.

Table 5. Comparisons with SOTA methods

Methods	Param.(M)	GFLOPS/%	C-View/%
Ta-CNN [4]	1.06	1.06	95.1
ST-GCN [2]	3.1	16.32	88.3
2s-AGCN [5]	6.94	37.32	95.1
4s-Shift-GCN [15]	2.76	10	96.5
SGN [9]	**0.69**	0.8	94.5
4s-Shift-GCN++ [16]	1.8	1.7	96.3
2s-ST-GCN++ [24]	1.39	2.8	96.7
EffiencientB4 [17]	1.1	4.05	96.1
MDKA-GCN(1s)	**0.69**	**0.69**	96.2
MDKA-GCN(2s)	1.38	1.38	96.6
MDKA-GCN(4s)	2.92	2.76	**96.8**

Comparison with Lightweight SOTA. To verify the overall performance of our model as shown in Table 5, we compare with SOTA methods in recent years in terms of accuracy model parameter volume and computational complexity. Compared to lightweight GCN method EfficientB4, our single-stream model has a smaller parameter size and computational complexity while achieving higher accuracy. We compare the model training process of MDKA-GCN and SGN on two datasets in Fig. 4. For fair comparison the hyperparameter settings and data preprocessing methods of the two models are kept consistent. Overall our method's overall performance has reached an advanced level and is more suitable for resource-limited mobile devices and practical application scenarios compared to most lightweight SOTA methods.

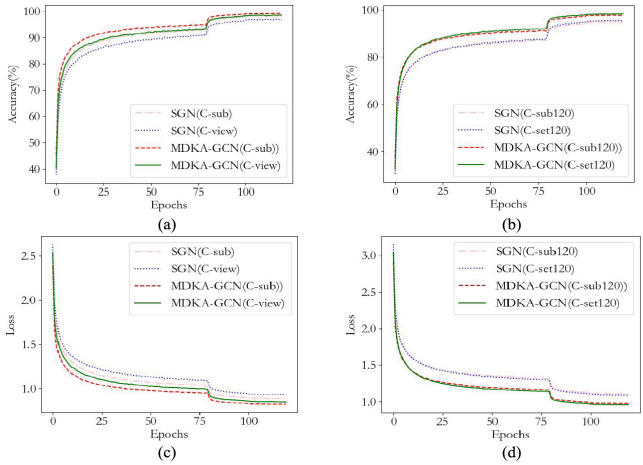

Fig. 4. comparison of the accuracy and convergence of MDKA-GCN and the baseline model SGN.

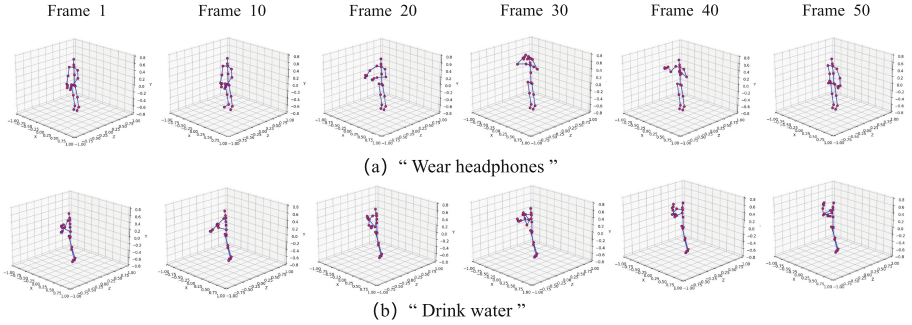

Fig. 5. Qualitative examples from NTU-RGB+D 60, Six frames are selected from each action.

5 Action Visualization

To more intuitively display the action process, this paper visualizes the skeleton diagrams of actions such as "wear headphones" and "drink water" by observing some similar or difficult-to-distinguish actions in Fig. 5, and selects a few frames from them. These actions are mainly completed by both hands and are extremely similar in spatial configuration and temporal dynamics, requiring long-term observation to distinguish.

6 Conclusion

In this paper, we introduce two semantic information into multiple graph convolution layers, and the model performance is improved while reducing

model parameters. Our designed branch convolution block emphasizes significant motion features, the proposed time multi-scale dilated convolution attention module enlarges the receptive field and enriches the representation ability of various temporal features. We conduct extensive experiments on current mainstream action recognition datasets, whose results show that MDKA-GCN is more effective than most mainstream methods in terms of computational cost and performance with broader application prospects in the future.

Acknowledgement. This work is supported by the National Natural Science Foundation of China (61976006).

References

1. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308 (2017)
2. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
3. Si, C., Chen, W., Wang, W., Wang, L., Tan, T.: An attention enhanced graph convolutional LSTM network for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1227–1236 (2019)
4. Xu, K., Ye, F., Zhong, Q., Xie, D.: Topology-aware convolutional neural network for efficient skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 2866–2874 (2022)
5. Shi, L., Zhang, Y., Cheng, J., Lu, H.: 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 (2019)
6. Qin, Z., et al.: Fusing higher-order features in graph neural networks for skeleton-based action recognition. IEEE Trans. Neural Netw. Learn. Syst., 1–15 (2022)
7. Wen, Y.H., Gao, L., Fu, H., Zhang, F.L., Xia, S., Liu, Y.J.: Motif-GCNS with local and non-local temporal blocks for skeleton-based action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **45**(2), 2009–2023 (2022)
8. Liu, Y., Zhang, H., Li, Y., He, K., Xu, D.: Skeleton-based human action recognition via large-kernel attention graph convolutional network. IEEE Trans. Visual Comput. Graphics **29**(5), 2575–2585 (2023)
9. Zhang, P., Lan, C., Zeng, W., Xing, J., Xue, J., Zheng, N.: Semantics-guided neural networks for efficient skeleton-based human action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1112–1121 (2020)
10. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
11. Wang, Y., Li, Y., Wang, G., Liu, X.: Multi-scale attention network for single image super-resolution. arXiv preprint arXiv:2209.14145 (2022)
12. Guo, M.H., Lu, C.Z., Liu, Z.N., Cheng, M.M., Hu, S.M.: Visual attention network. arXiv preprint arXiv:2202.09741 (2022)

13. Howard, A.G., et al.: MobileNets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
14. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1492–1500 (2017)
15. Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., Lu, H.: 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 (2020)
16. Cheng, K., Zhang, Y., He, X., Cheng, J., Lu, H.: Extremely lightweight skeleton-based action recognition with shiftGCN++. IEEE Trans. Image Process. **30**, 7333–7348 (2021)
17. Song, Y.F., Zhang, Z., Shan, C., Wang, L.: Constructing stronger and faster baselines for skeleton-based action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **45**(2), 1474–1488 (2022)
18. Zhang, H., Zu, K., Lu, J., Zou, Y., Meng, D.: EPSANet: an efficient pyramid split attention block on convolutional neural network. arxiv 2021. arXiv preprint arXiv:2105.14447 (2021)
19. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: 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 (2016)
20. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: NTU RGB+ D 120: a large-scale benchmark for 3d human activity understanding. IEEE Trans. Pattern Anal. Mach. Intell. **42**(10), 2684–2701 (2019)
21. Liu, Z., Zhang, H., Chen, Z., Wang, Z., Ouyang, W.: Disentangling and unifying graph convolutions for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 143–152 (2020)
22. Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., Tian, Q.: Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction. IEEE Trans. Pattern Anal. Mach. Intell. **44**(6), 3316–3333 (2021)
23. Miao, S., Hou, Y., Gao, Z., Xu, M., Li, W.: A central difference graph convolutional operator for skeleton-based action recognition. IEEE Trans. Circuits Syst. Video Technol. **32**(7), 4893–4899 (2021)
24. Duan, H., Wang, J., Chen, K., Lin, D.: PYSKL: towards good practices for skeleton action recognition. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 7351–7354 (2022)