Skeleton based human action recognition with relative position encoding

Shibin Xuan[a,b*], Kuan Wang[a], Li Wang[a], Chang Liu[a]
[a] School of Artificial Intelligence, Guangxi Minzu University, Nanning 530006, China; [b] Guangxi Key Laboratory of Hybrid Computation and IC Design Analysis, Nanning 530006, China

ABSTRACT

Aiming at the fact that the current graph convolution operation based on skeleton graph is limited in local adjacent nodes, or the overall relative position information of skeleton is omitted, an enhancement method of joint point position information based on relative position encoding of skeleton is proposed. The proposed method takes the central joint point of the human trunk as the root node, and all joint nodes form a tree structure according to the natural connection of the body, and the code of each joint node inherits the code of its parent node and also includes its own number in the sibling node. In addition, considering that the number of channels in the graph-based convolutional network model is generally larger, and the channel information itself has a strong correlation, the channel information frequency division and recombination operation is proposed to reflect the difference of information in different frequency bands in the channel. Experiments show that the proposed method has a certain effect on improving the effect of the embedded model.

Keywords: Action recognition, skeleton graph, relative position encoding

1. INTRODUCTION

Human behaviour recognition has always been one of the hot topics in computer vision. Due to the spatiotemporal features of human actions, human action representation or behaviours feature extraction is much more difficult than general images. Among them, the more typical ones are tracking method, optical flow method, RGB frame and skeleton method. Skeleton-based methods have received more attention from researchers in recent years. The main reason is that skeleton information can better reflect the changing characteristics of human limb action and is very simple, and it also avoids the influence of different clothing and environmental differences to the recognition results. The skeleton based human action are generally expressed as a vector sequence composed of three-dimensional coordinates of joint points, where the coordinates of joint points can be obtained by a pose estimation operator. Due to the many advantages of skeleton information, Skeleton-based behaviours recognition has attracted a lot of attention from researchers in recent years. Among them, skeleton-based behaviour recognition under the framework of deep learning has become the mainstream research direction. These methods fall broadly into the following categories: Recurrent Neural Network (RNN)[1], Convolutional Neural Network (CNN)[2], Graph Convolutional Network (GCN)[3]. The main function based on RNN is to reveal the dependence of video frames on the time axis, but its application scope is limited since it is not suitable for long-term memory. To improve the limitations of RNN, LSTM[4], GRU[5], and transformer[6] are used for behaviour recognition. In fact, these methods were originally used for natural language processing. Since both video sequences and natural language representations are in the form of sequences, these methods for processing natural language can also be directly used for video-based action recognition tasks. Liang[7] proposed a two-stream RNN to build a spatiotemporal model. Chunyu and Baochang[8] proposed a complex spatiotemporal model combining CNN and RNN with attention mechanism for skeleton-based human action recognition. Literature[9] utilizes multi-layer LSTMs to learn temporal dependencies in skeleton sequences. CNN-based recognition methods generally convert 3D skeleton sequences into a vector sequence. Wang[10] proposed the Joint Trajectory Maps (JTM) to represent spatial information and node trajectories. Carlos and Jessica[11] proposed to encode motion information by calculating the motion direction and size of skeleton link points. The GCN-based method regards the connection between joint points as a tree-like structure according to the structural features of the skeleton itself, and performs convolution operations on the established tree. Sijie and Yuanjun[3] first proposed a spatiotemporal graph convolutional network model (ST-GCN), which takes the joint points as the vertices of the graph, and connects the vertices in the graph with the natural connection features of the human body structure to form the edges of the graph to form the edges of the

* xuanshibin@gxmzu.edu.cn

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 1250613
 · 0277-786X · doi: 10.1117/12.2661803

graph. Graph convolution operations are performed on the graph, and general convolution operations are performed on the timeline. Maosen and Siheng[12] proposed the Action Structural Graph Convolutional Networks (AS-GCN), which establishes the dependencies between nodes by establishing activity chains and structural chains. Shi and Zhangd et al.[13] proposed the 2s-AGCN model, in which each GCN layer adaptively learns the graph structure according to the input graph skeleton data, and introduces the attention mechanism into the labeling of the importance of the edges of the structural graph. Lei Shi, Yifan Zhang et al.[14] proposed the MS-AAGCN model, which trains the model with 4 data streams, and uses a spatiotemporal attention model to enhance relevant nodes and features in important skeletons. Yang and Yan et al.[15] proposed the FGCN model, which gradually refined spatiotemporal features through a multi-step temporal sampling strategy and introduced a feedback mechanism in graph convolution. Liu and Zhang et al.[16] proposed MS-G3D to extract long-range feature information using multi-scale graph convolution techniques. The above improvement methods based on GCN are mostly at the cost of time and space. For this reason, people try to reduce the complexity of the algorithm without reducing the accuracy of the algorithm. Cheng and Zhang et al.[17, 18] proposed the ShiftGCN model and the ShiftGCN++ model, which replaced the graph and temporal convolution operations with the zero-FLOPs displacement graph operator, and introduced techniques such as knowledge distillation, spatial location encoding, and dynamic displacement graph convolution to improve model efficiency. Zhang and Lan et al.[19] proposed the SGN model, which uses the semantic information of joints and frames to design a compact semantically guided neural network to express the spatiotemporal correlation between joints and frames. Heidari and Iosifidis[20] proposed the TA-GCN model to select the key bones most conducive to activity recognition to perform spatiotemporal convolution operations on the skeleton sequence. In order to realize online human behavior recognition, Hedegaard and Heidari et al.[21] proposed Continual Spatio-Temporal Graph Convolutional Network (CoST-GCN), which reorganized the spatio-temporal graph convolutional neural network as a continuous inference network, which was processed without frame repetition. In this case, the step-by-step prediction function on the time axis is implemented. In addition, Duan and Zhao et al.[22] proposed the PoseConv3D model, which uses 3D heat-map volumes instead of graph sequences as human skeleton representations. The model is robust to noise and scene changes, and can handle multiple the human scene can also be easily integrated into other existing models. But the heat-map representation naturally uses more storage space than the general skeleton model. Song and Zhang[23] proposed an efficient Graph Convolutional Network (EfficientGCN) model, which further introduced SEpLayer, EpSepLayer and SGLayer layers based on ResGCN [24] to improve model efficiency.

The above methods pay more attention to the adjustment of the model structure, hoping that different network structures can obtain the implicit semantic features in the video skeleton sequence to improve the accuracy and efficiency of pattern recognition. This paper hopes to improve the recognition rate of the embedded model by enhancing the structural feature representation and changing the channel distribution structure without changing the existing model structure, because only a small number of parameters need to be learned when enhancing the representation of structural features, and this part The number of parameters relative to the number of parameters of the entire network model may or may not be counted. Therefore, the proposed method hardly increases the complexity of the original model.

2. RELEVANT RESEARCH

Considering that the proposed method does not involve modifying the model, it only improves the recognition accuracy of the existing model. To this end, in this paper, we mainly choose three most representative models as the embedding objects of our method, namely ST-GCN, MS-G3D and EfficientGCN models. These three models are briefly introduced as follows.

2.1 ST-GCN

The connection between the joints of the body conforms to the physiological structure of the human body, and the spatiotemporal graph model hopes to keep this connection relationship in the model. Suppose one frame in the skeleton sequence contains N joint points, and the entire sequence contains T frames, defining a space-time undirected graph$G = (V, E)$, where $V = \{u_{ti} | t = 1, \cdots, T; i = 1, \cdots, N\}$, $E = \{e_{st} | s, t = 1, \cdots, T \times N\}$, e_{st} represents the sth and tth vertex are two joint points that are physiologically connected in the same frame. The feature vector $F(u_{ti})$ at each vertex consists of the three-dimensional coordinate values of that point. It is assumed that the connection matrix A represents the correlation of each joint point in a frame, and the unit matrix I represents the self-correlation of each joint point in a frame. Then the graph convolution in one frame can be calculated as follows equation (1):

$$f_{out} = \Lambda^{-\frac{1}{2}}(A + I)\Lambda^{-\frac{1}{2}} f_{in} W \tag{1}$$

Where, f_{in} represents the input skeleton information. W denotes the channel weighted vector, Λ refers to a diagonal matrix, the element on its main diagonal $\Lambda^{ii} = \sum_j (A^{ij} + I^{ij})$, f_{out} refers to the output of a graph convolution operation. When it is divided into multiple subsets, the connection matrix is also decomposed into multiple sub-matrices, $A + I = \sum_j A_j$. Then equation (1) can be converted into equation (2):

$$f_{out} = \sum_j \Lambda_j^{-\frac{1}{2}} A_j \Lambda_j^{-\frac{1}{2}} f_{in} W_j \tag{2}$$

2.2 MS-G3D

MS-G3D believes that for robust skeleton-based action recognition methods, an ideal algorithm should have two characteristics: (1) Algorithms should go beyond local joint connectivity to extract multi-scale structural features and long-term dependencies, since structurally separated joints can also have strong correlations. (2) The algorithm is able to utilize complex joint relations across time and space for action recognition. To this end, the author proposes: (1) A multi-scale aggregation method, which eliminates redundant dependencies between joint nodes and can effectively capture joint relationships on human bones. (2) A new spatiotemporal graph convolution (G3D) model, which can promote the information to travel through space and time directly and effectively carry out feature learning. (3) The two methods mentioned above are combined to form a new feature extraction model (MS-G3D), which has the ability of multi-scale feature extraction across the spatial and temporal dimensions of the domain and further improves the performance of the model.

To this end, first the k-adjacency matrix is defined as equation (3).

$$[\tilde{A}_{(k)}]_{s,t} = \begin{cases} 1 & d(v_s, v_t) = k \\ 1 & s = t \\ 0 & \text{other} \end{cases} \tag{3}$$

Let $\widetilde{D}_{(k)}$is a diagonal matrix of $\tilde{A}_{(k)}$, Then GCN can be calculated according to the equation (4).

$$X_t^{(l+1)} = \sigma\left(\sum_{k=0}^{K} \widetilde{D}_{(k)}^{-\frac{1}{2}} \tilde{A}_{(k)} \widetilde{D}_{(k)}^{-\frac{1}{2}} X_t^{(l)} W_{(k)}^{(l)}\right) \tag{4}$$

Suppose the window on the time axis contains τ frames, and the block adjacency matrix is defined as equation (5).

$$\tilde{A}_{(\tau)} = \begin{bmatrix} \tilde{A} & \cdots & \tilde{A} \\ \vdots & \ddots & \vdots \\ \tilde{A} & \cdots & \tilde{A} \end{bmatrix} \in \mathbb{R}^{\tau N \times \tau N} \tag{5}$$

Then MS-G3D model can be defined as equation (6).

$$[X_{(\tau)}^{(l+1)}]_t = \sigma\left(\sum_{k=0}^{K} \widetilde{D}_{(\tau,k)}^{-\frac{1}{2}} \tilde{A}_{(\tau,k)} \widetilde{D}_{(\tau,k)}^{-\frac{1}{2}} [X_{(\tau)}^{(l)}]_t W_{(k)}^{(l)}\right) \tag{6}$$

2.3 EfficientGCN

To address the high complexity of GCN-based skeleton human recognition models severely limits its application and research, EfficientGCN first constructs multi-input branch (MIB) architecture that extracts the spatial structure and temporal motion features of the joints from the skeleton sequence. Second, four convolutional layers are added to CNN, namely, BottleLayer[25], SepLayer[26], EpSepLayer[27] and SGLayer[28]. Again, the composite scaling method[29] is used to obtain the structural hyper-parameters of each layer, which uses a set of fixed scaling coefficients to uniformly measure the width, depth and resolution of the network. Finally, the method in literature[30] was used to design the Spatial Temporal Joint attention module, which was inserted into each block of the model for more accurate recognition.

3. RELATIVE POSITION ENCODING AND CHANNEL FREQUENCY DIVISION STRATEGY

The spatiotemporal graph framework is built on the basis of fixed joint connections in the human body, which only determines that the nodes associated with it when the convolution of a certain point is executed participates in the calculation, and cannot reflect the overall joint position correlation. For this purpose, we propose to encode the nodes according to their relative relations, and integrate the relative position information into the coordinate information of the

node to further improve the spatial correlation characteristics of the skeleton. In addition, the contribution of the channel information of each joint point in the same frame to the corresponding posture change has a large difference. In order to reflect this difference, it is necessary to perform frequency division processing on the channel information.

3.1 Relative position encoding

To reflect the relative location relationship of each joint point, and inspired by the position encoding in the transformer, the relative position encoding is added to each joint point. This paper takes the skeleton of 25 joint points in the NTU-RGB-D database to introduce the relative position encoding of each joint point. Figure 1 shows the serial number of each joint point in the skeleton.

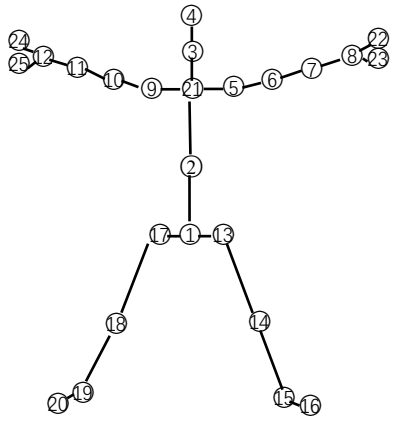

Figure 1. Human skeleton structure.

We take the human body center joint point numbered 2 as the root node, which is encoded as: [1, 0, 0, 0, 0, 0, 0], it has two children, numbered 1 and 21, and encode them respectively is [1, 1, 0, 0, 0, 0, 0] and [1, 2, 0, 0, 0, 0, 0]. The node numbered 21 has three children, whose numbers are 3, 5, 9, and the corresponding codes are: [1, 2, 1, 0, 0, 0, 0], [1, 2, 2, 0, 0, 0, 0] and [1, 2, 3, 0, 0, 0, 0]. And so on, we can get 7-bit integer codes for all 25 joint points. The specific codes are shown in Table 1.

Table 1. The corresponding code of each node.

No.	coding	No.	coding	No.	coding
1	[1, 1, 0, 0, 0, 0, 0]	10	[1, 2, 3, 1, 0, 0, 0]	19	[1, 1, 1, 1, 1, 0, 0]
2	[1, 0, 0, 0, 0, 0, 0]	11	[1, 2, 3, 1, 1, 0, 0]	20	[1, 1, 1, 1, 1, 1, 0]
3	[1, 2, 1, 0, 0, 0, 0]	12	[1, 2, 3, 1, 1, 1, 0]	21	[1, 2, 0, 0, 0, 0, 0]
4	[1, 2, 1, 1, 0, 0, 0]	13	[1, 1, 2, 0, 0, 0, 0]	22	[1, 2, 2, 1, 1, 1, 1]
5	[1, 2, 2, 0, 0, 0, 0]	14	[1, 1, 2, 1, 0, 0, 0]	23	[1, 2, 2, 1, 1, 1, 2]
6	[1, 2, 2, 1, 0, 0, 0]	15	[1, 1, 2, 1, 1, 0, 0],	24	[1, 2, 3, 1, 1, 1, 1]
7	[1, 2, 2, 1, 1, 0, 0]	16	[1, 1, 2, 1, 1, 1, 0]	25	[1, 2, 3, 1, 1, 1, 2]
8	[1, 2, 2, 1, 1, 1, 0]	17	[1, 1, 1, 0, 0, 0, 0],		
9	[1, 2, 3, 0, 0, 0, 0]	18	[1, 1, 1, 1, 0, 0, 0],		

It can be seen from Table 1 that the set of encoding numbers is {0, 1, 2, 3}. First, each number is mapped into a 5-dimensional vector using Embedding, and then a 1×1 convolution kernel is used to convert the 7 coding of each joint point

to a 3-dimensional vector, and this 3D vector is combined with the 3D coordinates of the joint points to form a 6-dimensional vector. Thus, the original 3-channel skeleton data is converted into 6-channel data information, and finally is sent to the GCN model.

3.2 Channel information frequency division and reorganization

In the GCN-based model, the input data gradually changes from a low number of channels to a high number of channels with an iterative process. As in the standard ST-GCN, the input is gradually changed from 3 channels to 64 channels, 128 channels and 256 channels. Especially in the case of high number of channels, the information dependence between channels is more obvious. To reflect this information dependency, and inspired by the different roles of high-frequency and low-frequency information in basic image enhancement, we use Fourier transform to divide the channels, which aims to separate high- and low-frequency information. Especially in the convolution operation on the time axis, the information of different frequency bands has different effects.

4. EXPERIMENTAL RESULTS

In order to verify whether the proposed method can effectively improve the effect of the existing human behavior recognition based on skeleton, two commonly used databases NTU RGB+D 60[31] and NTU RGB+D 120[32] are selected as test databases, and we select ST-GCN, MS-G3D and EfficientGCN as embedding models. For the sake of simplicity, the ablation experiments are only finished on the NTU RGB+D 60 database, and ST-GCN is chosen as the embedding model. The experimental system environment is Ubantu20.04.LTS and Cuda11.1, and the programming software is pytorch1.8.1.

4.1 Comparative experiment on the database NTU RGB+D 60

NTU RGB+D 60 contains 56,880 human activity videos, and consists of 60 categories. For the sake of simplicity, each video segment selects the first 300 frames at most, and the video with less than 300 frames is filled with 0 to 300 frames. Each frame contains 2 pieces of skeleton information, each skeleton contains 25 joint points, and the coordinate data of each joint point is composed of 3-dimensional vectors. Two recommended benchmarks:1) Xsub consists of 40320 training video segments and 16560 test sample sets, and the two datasets are from different individuals. 2) X-view is a dataset with video segments taken from cameras 2 and 3 as training samples, including 37,920 video segments, and video segments taken from camera 1 as a test samples, with a total of 18,960 video segments. Table 2 lists all the experimental results.

Table 2. Results on the database NTU RGB+D 60.

Methods	Xsub	X-view
ST-GCN	81.5	88.3
ST-GCN+our	87.62	92.76
MS-G3D	91.5	96.2
MS-G3D+our	91.92	96.35
EfficientGCN	92.1	96.1
EfficientGCN+our	92.22	96.15

According to the data in Table 2, the proposed method can improve the recognition rate of the embedded model without changing the structure of the original model, especially for the embedded model ST-GCN, the recognition rate is increased by 6.12% and 4.56%. The other two models are also improved to a certain extent, but the improvement effect is not obvious. Considering that the proposed method almost does not increase the parameters of the embedded model, and only increases a small part of the calculation amount, it is not enough for the existing high recognition rate. The effect of the original model has been improved, which is a very good result.

4.2 Comparative experiment on the database NTU RGB+D 120

NTU RGB+D 120 adds new video segments based on NTU RGB+D 60. It includes 114,480 video segments shot from 155 viewing angles by 106 individuals, and contains 120 activity categories. The two proposed training-test benchmarks

are: 1) X-sub120 includes 63026 videos for training and 50922 videos for evaluation. 2) X-set120 includes 54471 videos for training and 59477 videos are used to evaluate model performance. Table 3 lists all the experimental results.

Table 3. Results on the database NTU RGB+D 120.

Methods	X-sub120	X-set120
ST-GCN	70.7	73.2
ST-GCN+our	76.53	78.16
MS-G3D	86.9	88.4
MS-G3D+our	87.5	88.7
EfficientGCN	88.7	88.9
EfficientGCN+our	88.78	88.96

Table 3 shows that the proposed approach can still improve the recognition rate of the embedded model when the number of categories is higher, while for the basic ST-GCN model, the recognition rate is improved on the two test benchmarks respectively 5.83 and 4.96.

4.3 Ablation experiment

To reflect the influence of each part of the proposed method on the embedded model, also taking into account the complexity of the algorithm and the time used for the amount of data, and the purpose of ablation is to test the contribution of each part to the correct rate, we choose the simplest original ST- GCN model and NTU RGB+D 60 database. “Scattered error rate” in the literature[33] is also cited in the experiment, and Table 4 lists the experimental results.

Table 4. Ablation experiments on the database NTU RGB+D 60.

ST-GCN	Scattered error rate	Relative position encoding	Channel information frequency division	Recognition rate
√				81.5
√	√			85.27
√	√	√		86.10
√	√	√	√	87.62

The ablation results in Table 4 shows that the two strategies proposed in this paper have good effects on the original model, especially the proposed channel information frequency division and recombination operation is very simple to implement. It can be achieved by performing the channel Fourier transform only once, but from the experimental results, based on the previous option, the recognition rate can be improved by 1.52%, indicating that the information reorganization in the channel is of great research value.

5. CONCLUSION

Aiming at the situation that the relative position information between joint points in the skeleton graph is not fully reflected in the original skeleton graph network in the research of human action recognition, the relative position encoding of joint points is proposed. This encoding can not only ensure the local correlation without affecting the execution result of the graph convolution, but also take into account the correlation of the global joint points. This paper also analyses the influence of the current rare information distribution in the channel on the model, and proposes a channel information frequency division and recombination method. The experimental results show the effectiveness of the proposed approach. This is a problem that few people consider, and it is worth further research.

ACKNOWLEDGMENT

This research is supported by the National Natural Science Foundation of China (61866003).

REFERENCES

[1] Lev, G., Sadeh, G., Klein, B. and Wolf, L., "RNN fisher vectors for action recognition and image annotation," Eur. Conf. Comput. Vis., 833-850(2016).
[2] Chéron, G., Laptev, I. and Schmid, C., "P-CNN: Pose-based CNN features for action recognition," IEEE Int. Conf. Comput. Vis., 3218-3226(2015).
[3] Yan, S., Xiong, Y. and Lin, D., "Spatial temporal graph convolutional networks for skeleton-based action recognition," Proc. AAAI 32, 7444-7452(2018).
[4] Hochreiter, S, and Schmidhuber, J., "Long short-term memory," Neural Comput., 9(8),1735-1780(1997).
[5] Cho, K., Van Merriënboer, B., Bahdanau, D. and Bengio, Y., "On the properties of neural machine translation: Encoder-decoder approaches," arXiv preprint arXiv:1409.1259, (2014).
[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I., "Attention is all you need," Proc. 31st Int. Conf. on Neural Inf. Process. Syst., 6000-6010(2017).
[7] Wang, H. and Liang, W., "Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks," IEEE Conf. Comput. Vis. Pattern Recognit., 3633-3642(2017).
[8] Xie, C., Li, C., Zhang, B., Chen, C., and Liu, J., "Memory attention networks for skeleton-based action recognition," IEEE Trans. Neural Netw. Learn. Syst., doi: 10.1109/TNNLS.2021.3061115, 1-15(2021).
[9] Lin, L., Wu, Z., Zhang, Z., Yan, H. and Liang, W., "Skeleton-based relational modeling for action recognition," https://doi.org/10.48550/arXiv.1805.02556, (2018).
[10] Wang, P., Li, W., Li, C. and Hou, Y., "Action recognition based on joint trajectory maps with convolutional neural networks," Knowl.-Based Syst. 158(15), 43-53(2018).
[11] Caetano, C., Sena, J., Bremond, F., dos Santos, J. A. and Schwartz, W. R., "Skelemotion: A new representation of skeleton joint sequences based on motion information for 3d action recognition," IEEE Int. Conf. Adv. Vid. Sign.-based Surve., doi: 10.1109/AVSS.2019.8909840, 1-8(2019).
[12] Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y. and Tian, Q., "Actional structural graph convolutional networks for skeleton-based action recognition," IEEE Conf. Comput. Vis. Pattern Recognit., 3590-3598(2019).
[13] Shi, L., Zhang, Y., Cheng, J. and Lu, H., "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," IEEE Conf. Comput. Vis. Pattern Recognit., 12026-12035(2019).
[14] Shi, L., Zhang, Y., Cheng, J., Lu, H., "Skeleton-based action recognition with multi-stream adaptive graph convolutional networks," IEEE Trans. Image Process., 2020(29), 9532-9545(2020).
[15] Yang, H., Yan, D., Zhang, L., Sun, Y., Li, D. and Maybank, S. J., "Feedback graph convolutional network for skeleton-based action recognition," IEEE Trans. Image Process., 2022(31), 164-175(2022).
[16] Liu, Z., Zhang, H., Chen, Z., Wang, Z., and Ouyang, W., "Disentangling and unifying graph convolutions for skeleton-based action recognition," IEEE Conf. Comput. Vis. Pattern Recognit., 140-149(2020).
[17] Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J. and Lu, H., "Skeleton-based action recognition with shift graph convolutional network," IEEE Conf. Comput. Vis. Pattern Recognit., 183-192(2020).
[18] Cheng, K., Zhang, Y., He, X., Cheng, J. and Lu, H., "Extremely lightweight skeleton-based action recognition with shiftgcn++," IEEE Trans. Image Process., 2021(30), 7333-7348(2021).
[19] Zhang, P., Lan, C., Zeng, W., Xing, J., Xue, J. and Zheng, N., "Semantics-guided neural networks for efficient skeleton-based human action recognition," IEEE Conf. Comput. Vis. Pattern Recognit., 1109-1118(2020).
[20] Heidari, N. and Iosifidis, A., "Temporal attention-augmented graph convolutional network for efficient skeleton-based human action recognition," Proc. Int. Conf. Pattern Recognit., 7907-7914(2020).
[21] Hedegaard, L., Heidari, N. and Iosifidis, A., "Online skeleton-based action recognition with continual spatio-temporal graph convolutional Networks," arXiv:2203.11009v1, (2022).
[22] Duan, H., Zhao, Y., Chen, K., Lin, D. and Dai, B., "Revisiting skeleton-based action recognition," IEEE Conf. Comput. Vis. Pattern Recognit., 2969-2978(2022).
[23] Song, Y.-F. and Zhang, Z., "Constructing stronger and faster baselines for skeleton-based action recognition," IEEE Trans. Pattern Anal. Mach. Intell., doi: 10.1109/TPAMI.2022.3157033, (2022).
[24] Song, Y.-F. Zhang, Z., Shan, C. and Wang, L., "Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition," ACM Int. Conf. Multimedia, 1625-1633(2020).
[25] He, K., Zhang, X., Ren, S. and Sun, J., "Deep residual learning for image recognition," IEEE Conf. Comput. Vis. Pattern Recognit., 770-778(2016).
[26] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. and Adam, H., "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv:1704.04861, (2017).

[27] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L.-C., “Mobilenetv2: Inverted residuals and linear bottlenecks,” IEEE Conf. Comput. Vis. Pattern Recognit., 4510-4520(2018).
[28] Zhou, D., Hou, Q., Chen, Y., Feng, J. and Yan, S., “Rethinking bottleneck structure for efficient mobile network design,” Eur. Conf. Comput. Vis, 680-697(2020).
[29] Tan, M. and Le, Q. V., “Efficientnet: Rethinking model scaling for convolutional neural networks,” Int. Conf. Mach. Learn., 97, 6105-6114(2019).
[30] Hou, Q., Zhou, D. and Feng, J., “Coordinate attention for efficient mobile network design,” IEEE Conf. Comput. Vis. Pattern Recognit., 13713-13722(2021).
[31] Shahroudy, A., Liu, J., Ng, T.-T. and Wang, G., “NTU RGB+D: A large scale dataset for 3D human activity analysis,” IEEE Conf. Comput. Vis. Pattern Recognit., 1010-1019(2016).
[32] Liu, J., Shahroudy, A., Perez, M. L., Wang, G., Duan, L.-Y. and Chichung, A. K., “NTU RGB+D 120: A large-scale benchmark for 3d human activity understanding,” IEEE Trans. Pattern Anal. Mach. Intell., 42(10), 2684-2701(2020).
[33] Xuan, S., Wang, K., Liu, L., Liu, C. and Li, J., “Algebra based human skeleton sequence reduction and action recognition,” Proc. CECNet2021, 467-476(2021).