Graph Neural Networks for Skeleton-based action recognition

Kairen Chen[1], Zihao Yang[2,*] and Zhenyu Yang[2]

[1] Taiyuan University of Science and Technology, School of Computer Science, China

[2] Jiangxi University of Finance and Economics, School of Finance, China

*Yang@brsj010@qq.com

Kairen Chen and Zihao Yang are co-first author

Abstract. One of the important directions of the application of artificial intelligence based on human bone behavior recognition is also a research hotspot in the field of computer vision in recent years. Human image video not only contains complex backgrounds, but also uncertain factors such as changes in illumination and changes in the appearance of the human body, which makes behavior recognition based on image videos have certain limitations. Compared with image video, human skeleton video can well overcome the influence of these uncertain factors, so be- havior recognition based on human skeleton has received more and more attention. The human skeleton sequence not only contains the tempo- ral features, but also the spatial structure features of the human body. How to effectively extract the discriminative spatial and temporal fea- tures from the human skeleton sequence is a problem to be solved. In recent years, many methods have been applied to bone-based behavior recognition, such as Recurrent Neural Network (RNN), Convolutional Neural Network (CNN) and Graph Neural Network (GCN). This article will introduce the content and characteristics of these three methods one by one. , And conduct a comparative analysis on it.

Keywords: Human skeleton, action recognition, CNN, RNN, GCN.

1. ntroduction

Human action and behavior recognition is the most challenging research direc- tion in the field of computer vision, and is a current research hotspot. Its ultimate goal is to output the structural parameters of the whole or part of the human body, such as the outline of the human body, the position and orientation of the head, the position or part category of the human body joint. The research methods of behavior recognition almost cover all the theories and technologies in the field of computer vision, such as pattern recognition, machine learning, artificial intelligence, image graphics, statistics and so on. So far, many recog- nition methods have been proposed, and many important research results have been achieved. According to the different modalities of the input data, behavior recognition methods can be divided into the following two categories, namely:

RGB-based behavior recognition method and bone-based behavior recognition method. As shown in Figure 1, if the input is a natural continuous video frame by frame, that is, an image, use RGB-based methods, including traditional image processing methods, or deep neural network methods for classification, such as TSN, TSM; However, its shortcomings are also obvious. Changes in background, lighting, and appearance will have a great impact on it.For skeleton-based be- havior recognition method, as shown in Figure 2, the input is the bone data obtained by the pose estimation algorithm on the video, and the position of the joint node is represented by coordinate points. The node of the graph is the joint node, and the edge is the line between the joint nodes, forming a Skeleton graph. Then skeleton-based methods, such as ST-GCN and AGCN, are used to process the bone dot map, which can not be affected by background, lighting, appearance and other factors. Based on these advantages, bone data has attracted many people to conduct research in human action recognition, and the use of bone data can be expected to increase. With the continuous development of deep learning methods, many methods have been widely used in human skeleton-based behavior recognition and achieved very good results, such as recurrent neural network (RNN) and convolutional neural network Network (CNN), graph Neural network (GCN), etc. In this paper, we will expand and analyze these three methods, in- troduce their contents and characteristics

respectively, and then put these three methods together for comparison and analysis of their differences.

Fig. 1. World Map

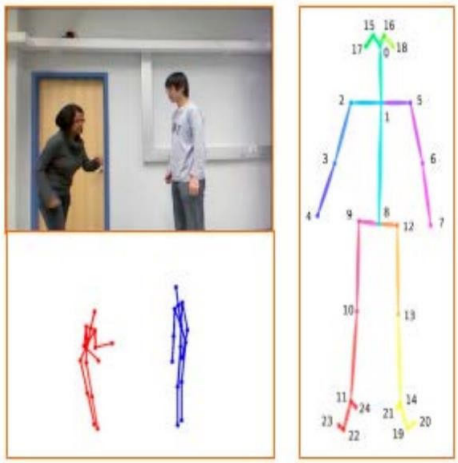


Fig. 2. Concrete and Constructions

2. Bone behavior recognition based on deep learning

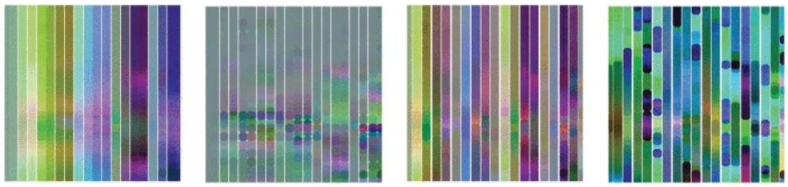
Existing studies have made quantitative and qualitative comparisons of existing behavior recognition techniques from RGB-based or skeleton-based perspectives, but not from the perspective of neural networks. To this end, we discuss and compare the CNN-based, RNN-based and GCN-based methods in detail. For each section, some recent related work will be introduced as a case based on some defect, such as a defect in one of the three models or a defect in the classical spatio-temporal modeling problem.

2.1 CNN

Convolutional neural networks have also been applied to skeleton-based action recognition. Different from RNN, CNN model can learn high-level semantic clues efficiently and easily, and it is endowed with excellent ability to extract high-level information. However, CNNs usually focus on image-based tasks, while action recognition tasks based on skeleton sequences will undoubtedly generate a serious time-dependent problem, so how to make more full use of 9spatial and temporal information in CNN-based architectures is still challenging. Usually, 3D skeleton sequence data are converted from vector sequences to pseudo-images in order to meet the needs of CNN input. However, it is usually not easy to represent information with both space and time, so many researchers encode bone joints as multiple two-dimensional pseudo-images and then input them into CNN to learn useful features [1, 2].

Wang [3] proposed Joint trajectory map (JTM), which represented the spatial configuration and dynamics of joint trajectory into three texture images through color coding. However, this approach is somewhat complicated and loses signifi- cance in the mapping process. To solve this shortcoming, Bo and Mingyi [4] used a translation-scale invariant image mapping strategy to first divide the human skeletal joints in each frame into five main parts according to the human body, and then map these parts to 2D form. This method makes the skeleton image contain both temporal and spatial information. Although the performance is improved, there is no reason to take the bone joint as an isolated point, because in the real world, there is a close connection between our bodies. For example, when you move your hands, consider not only the joints in your hands, but also other areas such as your shoulders and legs. Yanshan and Rongjie [5] proposed the shape motion representation of geometric algebra, which made full use of the information provided by skeletal sequences. Calos and Jessika [6] also pro- posed a new representation named Skel emotion based on motion information, which encodes temporal dynamics by explicitly calculating the amplitude and orientation values of

skeleton joints. Figure 3 shows the representation proposed by the shape motion [5], while Figure 4 illustrates the Skel emotion representation. In addition, similar to Skel emotion, [7] uses Skel emotion's framework but represents skeleton images based on tree structure and reference joints.



(a)Joint-Shape (b)Joint-Motion (c)Bone-Shape (d)Bone-Motion

Fig. 3. Single frame input

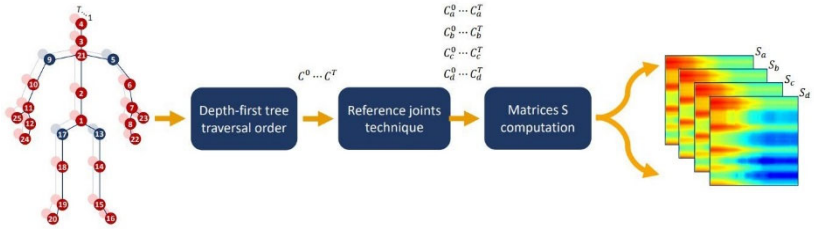


Fig. 4. The skeleton

Those CNN-based methods typically represent skeleton sequences as images by encoding temporal dynamics and skeleton joints as rows and columns, respectively, therefore, only if the convolution kernels can learn co-occurrence characteristic is considered, that is to say, some of the associated with the potential of all joints may be ignored, so CNN can't learn useful characteristics of the corresponding. Chao et al. used an end-to-end framework to learn collaborative features through a hierarchical approach, in which context information of different levels was gradually aggregated [8]. The point-level information is first encoded independently and then combined into a semantic representation of time-space. In addition to the representation of 3D skeleton sequence, CNN-based methods still have some other problems, such as model size and speed [9], CNN architecture (double-stream or three-stream [10]), occlusion, viewpoint change, etc. [9, 11]. So skeleton-based action recognition using CNN is still an open problem for researchers to explore.

2.2 RNN

In the main text [12], Duyong and Wangwei proposed a hierarchical RNN struc- ture, which divides the human skeleton into five parts according to the physical structure of the human body, and then inputs them into five sub-networks re- spectively, as shown in Figure 5. As the number of network layers increases, Representations extracted from subnets are hierarchically fused into higher-level input skeleton sequences and the final representation is fed into a single-layer perceptron for classification.

Literature [13]proposes a new dual-stream RNN architecture. It is used to model the temporal dynamics and spatial configuration of skeleton data. In the

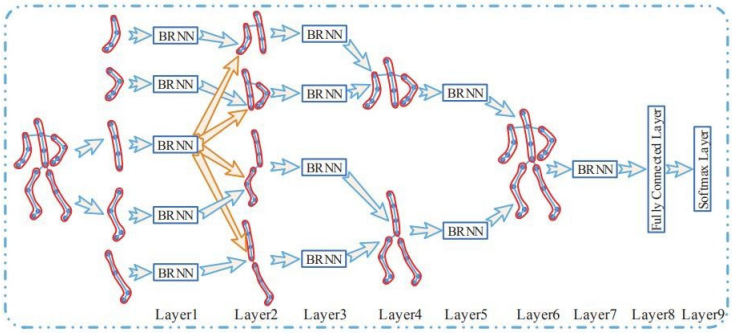


Fig. 5. Hierarchical RNN structure

data level preprocessing of spatial domain learning, the skeleton axis exchange model is used. Literature [14] uses the traversal method of a given skeleton se- quence to obtain the hidden relationship between two domains. Compared with the general method that places joints in a single chain and ignores the motion dependence between adjacent joints, the traversal method based on tree structure proposed in literature [14] will not add false connections when the relationship between joints is not strong enough. By using LSTM with trust gate to distinguish the input, if the input unit of tree structure is reliable, the storage unit is updated by inputting the latent spatial information. In reference [15], Chunyu and Baochang adopted attention RNN and CNN models, which facilitated complex spatiotemporal modeling. Firstly, a temporal attention module is intro- duced into the residual learning module to recalibrate the temporal attention of frames in the skeleton sequence. Then, a spatio-temporal convolution module is introduced into the first module to process the calibrated joint sequence as an image. Literature [16] adopts the attentional recurrent relation LSTM net- work, which uses the recurrent relation network to learn spatial features and multi-layer LSTM to learn the temporal features of skeleton sequences.

The problem of gradient explosion and disappearance of RNN-based structures is inevitable. LSTM and GRU may weaken these problems to a certain ex- tent, but they will become more prominent with the deepening of network layers. Literature [17] adds global context-aware attention to LSTM networks, which selectively focuses on information joints in bone sequences. Figure 6 shows the visualization effect of this method. Joints with more information are processed with red circles, indicating that these joints are more important for this special action.In addition, as data sets or depth sensors provide skeleton is not perfect, will influence the outcome of gesture recognition task, so the literature [17] con- verts skeleton to another coordinate system, in order to realize the robustness of scale, and then from the transformed data extract significant characteristics of the movement, rather than the original skeleton data sent to the LSTM.

In literature [18], Shuai and Wanqing proposed an independent recurrent neural network, which can solve the problem of gradient explosion and disappearance. Through it, a longer and deeper RNN can be built for high-semantic feature learning with stronger robustness. This improvement can be applied not only to skeleton-based action recognition, but also to language modeling and other fields. In this structure, a layer of neurons is independent of each other, so it can be used to process longer sequences.

Literature [19] proposed a new RNN structural representation, similar to the standard RNN, and proposed LSTM and GRU with gates and linear memory cells inside the RNN to compensate for the gradient and long time modeling problems of the standard RNN.

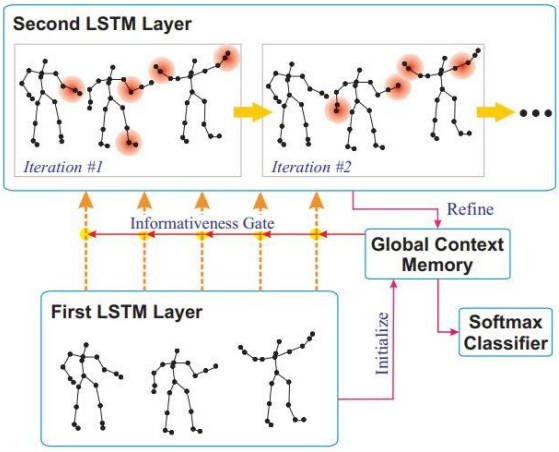


Fig. 6. The skeleton

The basic principle of action recognition is to make use of both temporal and spatial features. The structure of RNN allows to accept sequence data, so it is often used to deal with time series problems.

However, due to the lack of spatial modeling ability of RNN-based architecture, the performance of some related methods is difficult to obtain competitive results [20].

2.3 GCN

The human skeleton data is a natural topology map. The nodes in the topology map represent the joints of human bones, and the edges in the topology map represent the natural connections between bone joints. Therefore, the skeleton data itself is not a series of vectors or pseudo-image structures, but a kind of graphic structure data. Moreover, only from the perspective of bone, simply en- coding bone sequences as sequence vectors or 2D grids cannot fully express the dependence of related joints. Therefore, GCN, which can effectively represent graphical structure data, has recently been frequently used in skeleton behavior

recognition tasks. At present, there are two kinds of graph-related neural net- works: graph recurrent neural network (GNN) and graph convolutional neural network (GCN). This review focuses on GCN, and we will show some related advanced results. Graph convolutional neural networks (GCN), as a generaliza- tion form of CNN, can be applied to any structure including skeleton Graph. In GCN-based bone behavior recognition technology, the important problem is how to transform the original data into a specific graph structure.

Sijie and Yuanjun [21] first proposed a new model based on bone action recognition – Spatial Temporal Graph Convolution Networks (ST-GCN). In this network, human joints were first regarded as the vertices of the spatio-temporal graph. The edges that conform to the natural connection of joints are regarded as spatial edges, and the edges between the same joints in successive time steps are regarded as temporal edges. The input video is processed by the pose estimation algorithm to obtain the human bone point map, and then the bone point map is input to the constructed multi-layer spatio-temporal map for convolution, so that the information is aggregated along the spatial and temporal dimensions, and higher level feature maps are generated on the map. Finally, the classifica- tion of action categories is obtained by Softmax classifier. Figure 7 shows the constructed spatiotemporal map. This work has brought more attention to the advantages of using GCN for skeletal behavior recognition, and as a result, a number of related works have recently emerged.

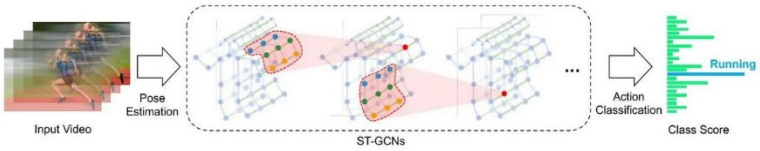


Fig. 7. The skeleton

The most common research focuses on the effective use of skeleton data [22, 23]. The motion structure graph convolutional network proposed by Maose and Siheng [22] can not only recognize human actions, but also output the next possible pose of the target using multi-task learning strategies. Two-stream Adaptive Graph Convolutional Networks (2S-AGCN) proposed by LeiShi and Yifan [23] can not only learn the connections between adjacent nodes, but also establish the connections between adjacent nodes. The fixed structure of skele- ton topology is solved.Figure structure of the work in constructing the adjacency matrix of two can learn B and C pectively study the common pattern in all the data and study alone in a single data model, and the manual to extract the key points of the second order information to capture the joint between the richer dependencies. Figure 8 shows the skeleton diagram structure of 2S-AGCN. The multi-parameter learning strategy used in this model is a good direction, because behavior recognition may be improved from other complementary tasks.

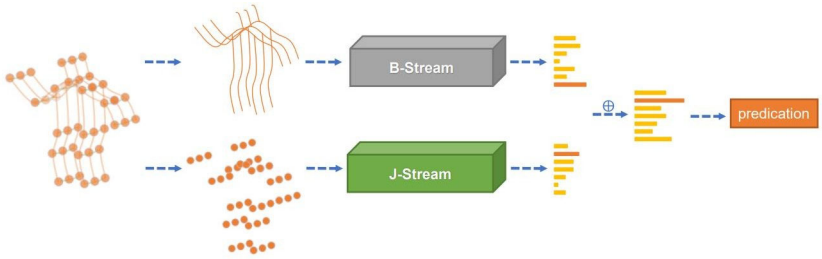


Fig. 8. The skeleton

According to the above introduction and discussion, the most attention is still data-driven, and what we need to do is to obtain the underlying information behind the 3D skeleton sequence data, while GCN-based behavior recognition mainly revolves around the ”how to get” problem, which is still an open and challenging problem. In particular, the skeletal data itself is spatio-temporal coupled, and the connection between joints and bones is also spatio-temporal coupled when the skeletal data is transformed into a graph.

3. Experimental data and methods

3.1 Introduction to Data Set

NTU RGB+D is a large-scale dataset collected and produced by Nanyang Tech- nological University for human action recognition, with 56,880 samples collected from 40 subjects in 60 categories of movements. Movements can be broadly di- vided into three categories: 40 everyday movements (such as drinking, eating and reading), 9 health-related movements (such as sneezing, stumbling and falling) and 11 interactive movements (such as punching, kicking and hugging).These actions are captured simultaneously by three Microsoft Kinect V2 cameras and contain multiple modal information, such as 3D skeleton joint position, depth, RGB frames, and infrared sequences.

3.2 The experimental method

The experimental results are shown in the following table1

4. Conclusion

In this paper, based on the behavior recognition of 3D skeleton sequence data, the three neural network methods CNN, RNN and GCN are systematically summa- rized and summarized, and the latest algorithms based on CNN, RNN and GCN technologies are introduced respectively. The advantages and disadvantages of

Table 1. Experimental methods and results

parameter	X-View(%)	X-sub(%)
CTR-GCN [24]	96.8	92.4
Shift-GCN [25]	96.5	90.7
DDGCN [26]	97.1	91.1
MS-G3D [27]	96.2	91.5
2s-AGCN [23]	95.1	88.5
AS-GCN [22]	94.2	86.8
ST-CGN [28]	88.3	81.5
CNN+Motion+Trans [4]	89.3	83.2
3scale ResNet152 [4]	92.3	85.0
Synthesized CNN [11]	87.2	80.0
Ind-RNN [18]	88.0	81.8
VA-LSTM [29]	87.7	79.2

ST-LSTM [30]	77.7	69.2
Deep-LSTM [31]	67.3	60.7

each method are analyzed and the problems under three different neural network structures are put forward.

The CNN method has some problems, such as model size and speed, CNN architecture, occlusion and viewpoint change. Rnn-based models are weak in spatial modeling, and the performance of some related methods is difficult to obtain competitive results. Behavior recognition based on GCN method needs to pay special attention to data-driven, considering how to obtain the underlying information behind 3D skeleton sequence data, which is a challenging problem. The application of GCN for behavior recognition is a very hot research direction in the future, and its superiority can be proved by comparing the accuracy of various methods in NTU-RGB + D dataset.

For the future development, if the application is to be more rapid in practice, a difficult problem that needs to be solved urgently is the recognition of real- time and long-term action behaviors. In addition, in the face of more complex data sets such as NTU-RGB + D 120, it is still a great challenge to improve the performance of the existing algorithms.

References

[1] Z. Ding, P. Wang, P. O. Ogunbona, and W. Li, “Investigation of different skeleton features for cnn-based 3d action recognition,” in 2017 IEEE International Confer- ence on Multimedia & Expo Workshops (ICMEW), pp. 617–622, IEEE, 2017.

[2] Y. Xu, J. Cheng, L. Wang, H. Xia, F. Liu, and D. Tao, “Ensemble one-dimensional convolution neural networks for skeleton-based action recognition,” IEEE Signal Processing Letters, vol. 25, no. 7, pp. 1044–1048, 2018.

[3] P. Wang, W. Li, C. Li, and Y. Hou, “Action recognition based on joint trajectory maps with convolutional neural networks,” Knowledge-Based Systems, vol. 158, pp. 43–53, 2018.

[4] B. Li, Y. Dai, X. Cheng, H. Chen, Y. Lin, and M. He, “Skeleton based action

[5] recognition using translation-scale invariant image mapping and multi-scale deep cnn,” in 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), pp. 601–604, IEEE, 2017.

[6] Y. Li, R. Xia, X. Liu, and Q. Huang, “Learning shape-motion representations from

[7] geometric algebra spatio-temporal model for skeleton-based action recognition,” in 2019 IEEE International Conference on Multimedia and Expo (ICME), pp. 1066– 1071, IEEE, 2019.

[8] C. Caetano, J. Sena, F. Br´emond, J. A. Dos Santos, and W. R. Schwartz, “Skelemo-

[9] tion: A new representation of skeleton joint sequences based on motion information for 3d action recognition,” in 2019 16th IEEE international conference on advanced video and signal based surveillance (AVSS), pp. 1–8, IEEE, 2019.

[10] C. Caetano, F. Br´emond, and W. R. Schwartz, “Skeleton image representation for

[11] 3d action recognition based on tree structure and reference joints,” in 2019 32nd SIBGRAPI conference on graphics, patterns and images (SIBGRAPI), pp. 16–23, IEEE, 2019.

[12] C. Li, Q. Zhong, D. Xie, and S. Pu, “Co-occurrence feature learning from skeleton

[13] data for action recognition and detection with hierarchical aggregation,” arXiv preprint arXiv:1804.06055, 2018.

[14] F. Yang, Y. Wu, S. Sakti, and S. Nakamura, “Make skeleton-based action recog-

[15] nition model smaller, faster and better,” in Proceedings of the ACM multimedia asia, pp. 1–6, 2019.

[16] A. Hernandez Ruiz, L. Porzi, S. Rota Bul`o, and F. Moreno-Noguer, “3d cnns on

[17] distance matrices for human action recognition,” in Proceedings of the 25th ACM international conference on Multimedia, pp. 1087–1095, 2017.

[18] M. Liu, H. Liu, and C. Chen, “Enhanced skeleton visualization for view invariant

[19] human action recognition," Pattern Recognition, vol. 68, pp. 346–362, 2017.

[20] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1110–1118, 2015.

[21] R. Zhao, H. Ali, and P. Van der Smagt, "Two-stream rnn/cnn for action recognition

[22] in 3d videos," in 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4260–4267, IEEE, 2017.

[23] H. Wang and L. Wang, "Modeling temporal dynamics and spatial configurations

[24] of actions using two-stream recurrent neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 499–508, 2017.

[25] C. Li, C. Xie, B. Zhang, J. Han, X. Zhen, and J. Chen, "Memory attention networks

[26] for skeleton-based action recognition," IEEE Transactions on Neural Networks and Learning Systems, 2021.

[27] L. Li, W. Zheng, Z. Zhang, Y. Huang, and L. Wang, "Skeleton-based relational

[28] modeling for action recognition," arXiv preprint arXiv:1805.02556, vol. 1, no. 2, p. 3, 2018.

[29] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot, "Global context-aware atten-

[30] tion lstm networks for 3d action recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1647–1656, 2017.

[31] S. Li, W. Li, C. Cook, C. Zhu, and Y. Gao, "Independently recurrent neural net-

[32] work (indrnn): Building a longer and deeper rnn," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5457–5466, 2018.

[33]

[34] P. Zhang, J. Xue, C. Lan, W. Zeng, Z. Gao, and N. Zheng, "Adding attentive- ness to the neurons in recurrent neural networks," in proceedings of the European conference on computer vision (ECCV), pp. 135–151, 2018.

[35] D. Wu and L. Shao, "Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 724–731, 2014.

[36] 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, 2018.

[37] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3595– 3603, 2019.

[38] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "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.

[39] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, "Channel-wise topology refinement graph convolution for skeleton-based action recognition," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13359–13368, 2021.

[40] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "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.

[41] M. Korban and X. Li, "Ddgcn: A dynamic directed graph convolutional network for action recognition," in European Conference on Computer Vision, pp. 761–776, Springer, 2020.

[42] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "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.

[43] J. Gesnouin, S. Pechberti, G. Bresson, B. Stanciulescu, and F. Moutarde, "Predict- ing intentions of pedestrians from 2d skeletal pose sequences with a representation- focused multi-branch deep learning network," Algorithms, vol. 13, no. 12, p. 331, 2020.

[44] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, “View adaptive recur- rent neural networks for high performance human action recognition from skeleton data,” in Proceedings of the IEEE international conference on computer vision, pp. 2117–2126, 2017.

[45] J. Liu, A. Shahroudy, D. Xu, and G. Wang, “Spatio-temporal lstm with trust gates for 3d human action recognition,” in European conference on computer vision, pp. 816–833, Springer, 2016.

[46] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, “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.