UTM UNIVERSITI TEKNOLOGI MALAYSIA
INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

A Review: Deep Learning for 3D Reconstruction of Human Motion Detection

Junzi Yang, Ajune Wanis Ismail
Mixed and Virtual Reality Research Lab, Vicubelab
School of Computing
Universiti Teknologi Malaysia
81310 UTM Johor Bahru, Johor, Malaysia
Email: yangjunzi@graduate.utm.my, ajune@utm.my

Submitted: 26/9/2021. Revised edition: 27/11/2021. Accepted: 12/12/2021. Published online: 16/5/2022
DOI: https://doi.org/10.11113/ijic.v12n1.353

***Abstract*—3D reconstruction of human motion is an important research topic in VR/AR content creation, virtual fitting, human-computer interaction and other fields. Deep learning theory has made important achievements in human motion detection, recognition, tracking and other aspects, and human motion detection and recognition is an important link in 3D reconstruction. In this paper, the deep learning algorithms in recent years, mainly used for human motion detection and recognition, are reviewed, and the existing methods are divided into three types: CNN-based, RNN-based and GNN-based. At the same time, the main stream data sets and frameworks adopted in the references are summarized. The content of this paper provides some references for the research of 3D reconstruction of human motion.**

***Keywords*—3D Reconstruction, Human Motion, Deep Learning**

I. INTRODUCTION

The wave of artificial intelligence has swept the world, computer vision technology as an important part of artificial intelligence has made great progress. In particular, the rise of deep learning algorithms has injected endless vitality into computer vision technology. As a multi-layer perceptron used to recognize two-dimensional images, Convolutional Neural Network has strong learning ability, which is similar to biological neural network. It can learn image features independently and reduce the complexity of network model in computer vision by weighting network structure. Therefore, computers have gradually made amazing achievements in such tasks as classification, segmentation, positioning and detection. Its related applications are also beginning to develop, affecting our lives more profoundly. At the same time, the research of 3D reconstruction has also opened a new page. Human movement is one of the most common visual content in the real and virtual worlds. Since almost all human behavior understanding requires accurate motion reconstruction, human motion reconstruction has become a hot topic in machine vision research. The 3D reconstruction of human motion is to reconstruct the key joints or parts of the human body from the given graphics, videos and sensor information, and then combine them into a whole to achieve the effect of motion reconstruction on the basis of the hierarchical connection relationship of human joints. At present, the 3D reconstruction technology of human motion has been widely applied in human-computer interaction, costume design, virtual fitting, VR games.

The earliest study of human movements was proposed by Johansson [1] in 1973. The motion of the main joints is described by a few bright spots against a dark background. The effectiveness of the kinetic-geometric model for visual vector analysis is verified. This study opens the door for human motion analysis. Since then, a number of research methods have emerged. At present, deep learning is the most popular method for 3D reconstruction of human motion detection, and different algorithms have different processing methods. Based on the standard RNN model typically used for human motion, Martinezet *et al*. [2] develop a sequence-sequence model with residual connections. Its performance is better than the early human motion prediction work and achieves good results. Li *et al*. [3] propose a dynamic multi-scale graph neural network (DMGNN) which is adaptive during training and a multiscale graph computational unit (MGCU). Despite the continuous improvement of deep learning algorithms, motion pose estimation has been a recognized problem for researchers in the study of computer

vision. A good attitude estimation method needs to be robust to occlusion and deformation, stable to changes caused by factors such as illumination and clothing, and the human body is a hinged object with different attitudes, so it is difficult to keep absolute static. In this paper, the most advanced deep learning algorithms for 3D reconstruction technology of human motion detection are reviewed to provide reference for further research. Firstly, the foundation of deep learning algorithms is introduced in detail. Then, according to different depth learning algorithms, the latest research progress is introduced from three aspects, and the existing problems of the algorithm are discussed. Finally, the framework of deep learning and the data set used in research literature are introduced.

II. Deep Learning Fundamentals

For different types of data and problems, people have studied all kinds of neural network structural models. Now, the mainstream methods of deep learning technology in 3D reconstruction of human motion detection are mainly CNN, RNN and GNN as the basic framework or their combination, and have achieved remarkable results.

Fig. 1 shows the overall structure of convolutional neural network, which is mainly composed of convolution layer, activation function, pooling layer and full connection layer. LeCun *et al*. first proposed Convolutional Neural Network (CNN) in 1998 [4]. The structure of LeNet network is divided into eight layers, as shown in Fig. 1, which mainly uses the principle of image local correlation to process image data. AlexNet is a Convolutional Neural Network developed by Krizhevsky *et al*. In the ILSVRC competition of that year, the error rate of Top5 was reduced to 15.315%. Compared with Lenet-5, AlexNet uses a deeper network structure, with 5 convolution layers, 3 full connection layers, 60 million parameters and 65,000 neurons. AlexNet uses two Gpus for calculation, which greatly improves computing efficiency. The sigmoid function and Tanh function are replaced with the non-saturated nonlinear function ReLU function [5]. Since AlexNet, deep learning has enjoyed the Renaissance. In the following years, various Convolutional Neural Network models based on the basic structure of AlexNet spring up, such as VGGNet, GooleNet and ResNet[6]. The proposal of ResNet is an improvement on the degradation of deep network structure, which is a milestone event in the history of CNN image processing. Since then, the research focus of the academia has changed from how to improve the accuracy of the neural network to how to achieve the same accuracy with less parameters and calculation. SqueezeNet is a typical example [7]. With the deepening of deep learning research, more CNN models and design ideas have been adopted into the network model design of 3D human motion reconstruction, which greatly promotes the development of 3D human motion reconstruction technology.

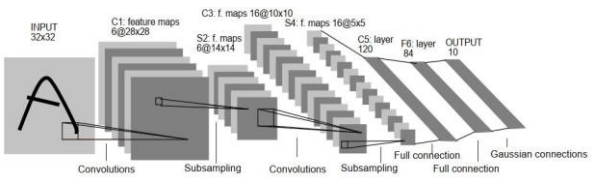

Fig.1. Schematic diagram of CNN [4]

RNN (Recurrent Neural Network) is used to process sequence data. The difference between CNN and RNN is that a directional loop is formed between neurons, in which the hidden state at the last moment and the input at this moment are both the input of neurons, so the network can remember the information at the previous moment. The structure of the circulation unit is shown in Fig. 2. RNN is unidirectional propagation. Based on this, bidirectional RNN [8] is proposed. In the process of training RNN, due to the repeated multiplication of weights, the problems of gradient disappearance and explosion are easy to occur, and it is difficult to learn for a long time. Therefore, RNN can only deal with the problems of short sequence dependence. LSTM [9] is an improvement on RNN. Unlike the cyclic layer in the basic structure of RNN, LSTM uses a gate control mechanism in the memory unit and combines short-term memory with long-term memory. It can learn the content with long time dependence and alleviate the problem of gradient explosion and disappearance to a certain extent. GRU is improved on the basis of LSTM. It has the same effect as LSTM, but it is improved in structure. GRU [10] simplifies the three "gates" of LSTM structure to two "gates". To prevent the gradient from disappearing and exploding, IndRNN [11] introduces Relu as the activation function, and separates the neurons in the layer, which can also build a deeper and longer network and make the network learn for a long time. Dual-path Recursive Neural Network (DPRNN) [12] is an effective and simple way to organize RNN layers in deep structures to make RNN model long sequences. The experimental result shows that replacing one-dimensional CNN with DPRNN in TasNet can improve the experimental results.

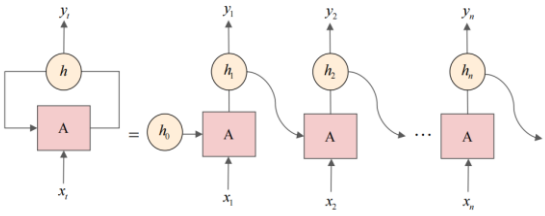

Fig.2. Schematic diagram of RNN

Scarselli *et al.* first proposed the concept of graph neural network in their paper [13]. In the paper, they used neural networks on graph structure data. GNN also has many limitations and is suitable for shallow structures, most with

no more than three layers. Graph Convolutional Neural Network (GCN) summarizes the convolution operation from grid data to graph data, which is a combination of CNN and graph topology structure, and implements multi-layer stacking. When constructing GCN, two methods, spectral method and non-spectral method, are usually followed. The Graph Attention Network (GAT) [14] introduces the attention mechanism based on GCN, and its architecture is shown in Fig. 3. Masked self-attentional layers are introduced to improve the disadvantages of graph convolution. Assigning corresponding weights to different adjacent nodes requires neither matrix operation nor prior knowledge of the graph structure. The model has better performance and is more robust to disturbances. Relational Graph Attention Networks [15] applies the Attention mechanism to graph convolution and adds relational information to the model, thus extending the non-relational graph attention mechanism. Relational Graph Attention Networks is an extension of GAT and has broader applications. Aiming at the problems of GCNs, Self-Supervised Semantic Alignment for Graph Convolution Network (SelfSAGCN) is proposed [16]. Identity Aggregation and Semantic Alignment are its two key approaches. This algorithm reduces over-smoothing and enhances the similarity between unlabeled features and labeled features of the same class. Experimental results show that the algorithm is better than other methods in various classification tasks.

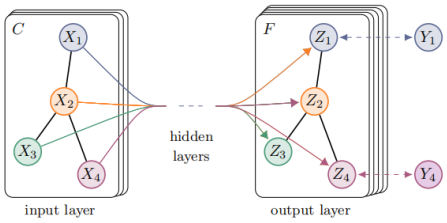

Fig.3. Graph Convolutional Neural Network [14]

III. Deep Learning Algorithms for 3D Reconstruction of Human Motion

3D reconstruction of human motion detection is a very complicated process. As discussed earlier, the mainstream approaches to deep learning technologies have their own characteristics. Depending on the specific properties of these methods, how to apply these deep learning methods to 3D reconstruction of human motion is extremely important. Based on the types of deep learning methods, this section elaborates 3D reconstruction of human motion detection from the following three aspects: (I) CNN based, (Ⅱ) RNN based, (Ⅲ) GNN based human motion reconstruction.

1. CNN-based Approach

CNN is the most widely used neural network among all the neural networks, and it is also the first one used for three-dimensional reconstruction of human motion detection. It can process images and any kind of data that can be converted into a similar image structure. Tompson *et al.* [17] propose to use CNN to make pose estimation and use heatmap to regression the key points. Fig. 4 illustrates the main flow. The method optimizes the prediction results by using the structural relations between key points and markov random field. LeCun's team proposes a novel architecture in which refined models are covered with the latest CNN models, including an effective "position refinement" model that can be trained to estimate joint offset positions in small areas of the image [18]. Stacked hourglass networks are a cascade of funnel-like neural networks, each of which acts as an encoder and decoder to extract features and generate heat map results [19]. In recent years, many studies on human pose estimation (single or multiple) have been based on this basic network structure, as well as another network structure, OpenPose. OpenPose is an open source project at Carnegie Mellon University based on models from three papers. One of the papers describes the 2D pose detection method PAF (Part Affinity Fields) in multi-player images, which first detects various points and then connects them with individuals to realize real-time detection of multiple people [20]. Ke *et al.* [21] improve the recent deep convolution and deconvolution hourglass model in four key points, and develop a robust multi-scale structural perceptual neural network for human pose estimation.

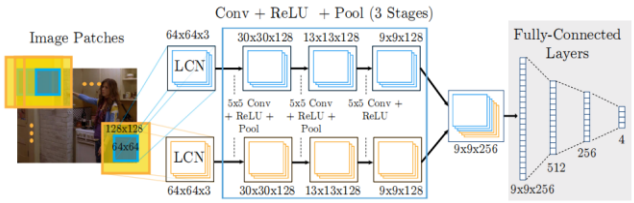

Fig.4. Multi-Resolution Sliding-Window With Overlapping Receptive Fields [17]

2. RNN-based Approach

The dynamic characteristics and context-dependent information of action can be captured by using recurrent neural network. Based on layered bone input, a multi-layer RNN framework is proposed by Du *et al.* [22]. The diagram of the proposed hierarchical recursive neural network is shown in Fig. 5. In the algorithm, the human body is divided into five parts, and then each part is input into five subnets for training. In the end, the extracted features are input into a single layer perceptron to determine the action category. Based on the Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM), an attentional mechanism is proposed to learn temporal and spatial features of skeleton data in human motion recognition tasks [23]. In the attention subnetwork of spatial dimension, the author uses LSTM network to learn the relationship between the nodes of the current frame and the nodes of the previous frame, form a currently input attention map frame node data, and

automatically find the current frame data of skeleton points, which has the greatest impact on action recognition. In the attention subnetwork of the time dimension, the author uses the LSTM network to learn the relationship between the current frame and the previous frame, to form the attention map of the current input frame data, and automatically learn which video frames contribute the most to action recognition. Zhang *et al.* [24] propose an adaptive recursive neural network (RNN) based on LSTM structure, instead of relocating the skeleton based on human defined prior criteria. This allows the network itself to adapt from one end to the other to the most appropriate point of view.

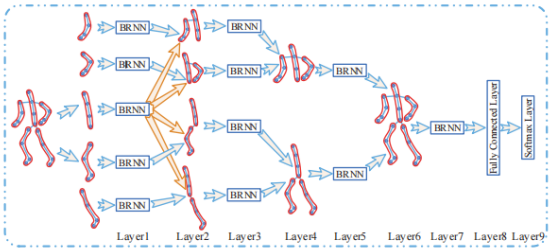

Fig.5. An illustrative sketch of the proposed hierarchical recurrent neural network [22]

Multistage convolutional neural network (CNN) has made advanced achievements in realizing single image human posture estimation, but its application to video requires a lot of calculation, and there will be performance degradation and jitter. A new recursive LSTM model is proposed for video pose estimation [25]. Artacho & Savakis [26] proposed UniPose and Unipos-LSTM architectures for single image and video pose estimation respectively. This structure can better understand the context information in the framework and help to estimate the posture of the subject more accurately.

3. GNN-based Approach

GNN can effectively extract important features from graph data and regard human skeletal connections as undirected graphs, where joint nodes are graph nodes and joint edges are graph edges. Yan *et al.* [27] propose to construct a human skeleton map into an undirected weighted spatiotemporal map network, and applied graph convolutional neural network in human action recognition for the first time, as shown in Fig. 6. The typical part-based method divides the human skeleton into multiple subgraphs, and defines the information transmission mode within and between subgraphs, so that local joint information can be extracted better and information transmission speed can be accelerated at the same time [28]. The double-flow adaptive graph convolution network proposed by Shi *et al.* [29] can adaptively learn the dynamic connecting edges between bone points, which improves the flexibility of the model. Meanwhile, this model proposes a dual-flow framework for modeling the first-order and second-order information simultaneously, which significantly improves the accuracy of recognition. Obinata and Yamamoto [30] develop a module based on ST-GCN to add connections between adjacent nodes between frames. The co-occurrence feature in the time domain is captured with good results. Symbiotic Graph Neural Networks [31] fuses many methods described above, and proposes a training method of mutual supervision between prediction and classification tasks. The ability of the model to obtain joint features is effectively improved.

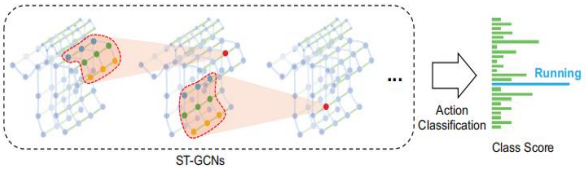

Fig.6. Action classification based on spatio-temporal graph convolutional neural network[27]

When collecting image data, data is often lost due to weather, light, occlusion and other reasons. To solve the problems of incomplete motion data and difficult human motion prediction, a new multi-task graph convolution network (MTGCN) [32] is proposed. Instead of dealing with them individually, repairing missing values in the observed sequence and predicting human behavior, the approach considers the two supervisory tasks jointly.

4. Discussion of Algorithms

In the actual scene, 3D reconstruction of human movement is a very complex process, which requires human movement detection and recognition. Although deep learning algorithms have made great achievements in this aspect, there are still many problems in the above algorithms, including the following aspects.

(1) The complexity of many algorithms is high. The model still needs a large number of parameters, and even the most advanced behavior recognition algorithms rely on the top of the central processing unit (CPU) and graphics processing unit (GPU) to provide computing power support, which consumes memory and has a slow speed in the process of operation. Ordinary computers cannot run relevant algorithms, and large-scale promotion and deployment cannot be achieved, so it is not easy to be applied in practice. The knowledge distillation method uses the knowledge learned from the large model network with good performance to train the network with few parameters and weak learning ability, so as to achieve the effect of model compression. Li Z *et al.* [33] have started to use knowledge distillation to estimate human posture. This is going to be a research trend. In addition, the lightweight model SqueezeNet has been successfully applied in the field of image classification, so the idea of depth separable convolution or grouping convolution can be applied in the

design of sequential feature extraction module to reduce the parameters of the model.

(2) Most deep learning models require supervised learning to extract human action features, so a large number of labeled data sets are required. Moreover, some models are suitable for human action recognition in specific scenes, and it is difficult to achieve generalization. In the new environment, it is necessary to re-collect data for processing and annotation, which requires a lot of time and energy. At present, transfer learning algorithms have made a large number of achievements. Applying the transfer learning algorithm to human motion detection and recognition can effectively solve the problem of small sample size and improve the generalization ability of the model. At the same time, some achievements have also been made in small sample learning. For example, Xian Y *et al*. [34] propose a two-stage method, first learning video features on the base class, and then fine-tuning the classifier on the new class.

(3) Human movements are ever-changing and easily affected by light changes, noise and occlusion, so feature extraction is difficult, and more fine-grained recognition of human movements is needed. At present, the performance of deep learning algorithms in fine-grained action recognition has not reached a high level. Weak supervision can be used to locate the key parts that can distinguish the action categories, and the distinguishing line features can be used as auxiliary to complete the classification.

IV. Deep Learning Experiment for 3D Reconstruction of Human Motion Detection

With the improvement of computer hardware, various open source deep learning frameworks are emerging one after another, among which the mainstream deep learning frameworks mainly include TensorFlow, Torch, Caffe, Theano, Keras, etc. In the 3D reconstruction of human motion detection based on deep learning, data sets are also essential to ensure the accuracy of the method and the consistency of evaluation. The data sets are mainly from public data sets. Table Ⅰ shows the data sets and frameworks used in the research work of 3D reconstruction of human motion detection.

TABLE I. 3D Reconstruction Dateset and Framework

Reference	DateSet	Method	Framework
Tompson *et al*., 2014	FLIC; LSP	CNN+Graphical Model	Torch7
Tompson *et al*., 2015	FLIC; MPII-Human-Pose	CNN	Torch7
Newell *et al*., 2016	FLIC; MPII-Human-Pose	Stacked Hourglass Networks	Torch7
Cao *et al*., 2017	MPII-Human-Pose; COCO human pose	PAFs	–
Ke *et al*., 2018	FLIC; MPII-Human-Pose	multi-scale structure-aware neural network	–
Du *et al*., 2015	MSR Action3D; HDM05	HBRNN-L	–
Song *et al*., 2017	the SBU Kinect interaction; NTU RGB+D	STA-LSTM	–
Zhang *et al*., 2017	the SBU Kinect interaction; NTU RGB+D; SYSU 3D Human-Object Interaction Set	VA-LSTM	Theano
Luo *et al*., 2018	Penn Action ; Sub-JHMDB	LSTM PM	Caffe
Artacho and Savakis, 2020	MPII; Leeds Sports Pose; Penn Action; BBC Pose	UniPose	PyTorch 1.0
Yan *et al*., 2018	Kinetics ; NTU-RGBD	ST-GCN	Pytorch
Thakkar and Narayanan, 2018	NTURGB+D; HDM05	PB-GCN	Pytorch
Shi *et al*., 2019	NTU-RGBD; Kinetics	2s-AGCN	PyTorch
Obinata and Yamamoto, 2021	NTU RGB+D; Kinetics-Skeleton	MS-AAGCN+TEM	PyTorch
Li *et al*., 2021	NTU-RGB+D; Kinetics; Human3.6M; CMU Mocap	Sybio-GNN	PyTorch 0.4.1.
Cui and Sun, 2021	H3.6M; CMU MoCap; 3DPW MoCap	MT-GCN	–

V. Conclusion

With the improvement of material living standards, the aggravation of environmental pollution and global aging, people have put forward higher development requirements for medical and health care, intelligent equipment, green energy and other fields. It will certainly improve people's quality of life and happiness if we deeply study the laws of human movement and apply them to fields closely related to human life. This paper introduces some deep learning methods for 3D reconstruction of human motion detection, but there are still many problems to be solved.

ACKNOWLEDGMENT

We are grateful to Mixed and Virtual Reality Laboratory (mivielab) in Vicubelab at Universiti Teknologi Malaysia (UTM). This work was funded by UTM-GUP Funding Research Grants Scheme (Q.J130000.2628.14J85).

References

[1] Johansson, G. (1973). Visual Perception of Biological Motion and a Model for Its Analysis. *Perception & Psychophysics*, 14(2), 201-211.

[2] Martinez, J., Black, M. J., & Romero, J. (2017). On Human Motion Prediction Using Recurrent Neural Networks.

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2891-2900.

[3] Li, M., Chen, S., Zhao, Y., Zhang, Y., Wang, Y., & Tian, Q. (2020). Dynamic Multiscale Graph Neural Networks for 3d Skeleton Based Human Motion Prediction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 214-223.

[4] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11), 2278-2324.

[5] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 25, 1097-1105.

[6] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770-778.

[7] Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. arXiv preprint arXiv:1602.07360.

[8] Schuster, M., & Paliwal, K. K. (1997). Bidirectional Recurrent Neural Networks. *IEEE Transactions on Signal Processing*, 45(11), 2673-2681.

[9] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-term Memory. *Neural Computation*, 9(8), 1735-1780.

[10] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-decoder for Statistical Machine Translation. arXiv preprint arXiv:1406.1078.

[11] Li, S., Li, W., Cook, C., Zhu, C., & Gao, Y. (2018). Independently Recurrent Neural Network (indrnn): Building a Longer and Deeper rnn. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5457-5466.

[12] Luo, Y., Chen, Z., & Yoshioka, T. (2020, May). Dual-path rnn: Efficient Long Sequence Modeling for Time-domain Single-channel Speech Separation. *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE*, 46-50.

[13] Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2008). The Graph Neural Network Model. *IEEE Transactions on Neural Networks*, 20(1), 61-80.

[14] Kipf, T. N., & Welling, M. (2016). Semi-supervised Classification with Graph Convolutional Networks. arXiv preprint arXiv:1609.02907.

[15] Busbridge, D., Sherburn, D., Cavallo, P., & Hammerla, N. Y. (2019). Relational Graph Attention Networks. arXiv preprint arXiv:1904.05811.

[16] Yang, X., Deng, C., Dang, Z., Wei, K., & Yan, J. (2021). SelfSAGCN: Self-supervised Semantic Alignment for Graph Convolution Network. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16775-16784.

[17] Tompson, J. J., Jain, A., LeCun, Y., & Bregler, C. (2014). Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation. *Advances in Neural Information Processing Systems*, 27, 1799-1807.

[18] Tompson, J., Goroshin, R., Jain, A., LeCun, Y., & Bregler, C. (2015). Efficient Object Localization using Convolutional Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 648-656.

[19] Newell, A., Yang, K., & Deng, J. (2016). Stacked Hourglass Networks for Human Pose Estimation. *European Conference on Computer Vision*, Springer, Cham. 483-499.

[20] Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. (2017). Realtime Multi-person 2d Pose Estimation Using Part Affinity Fields. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7291-7299.

[21] Ke, L., Chang, M. C., Qi, H., & Lyu, S. (2018). Multi-scale Structure-aware Network for Human Pose Estimation. Proceedings of the European Conference on Computer Vision (ECCV), 713-728.

[22] Du, Y., Wang, W., & Wang, L. (2015). Hierarchical Recurrent Neural Network for Skeleton based Action Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1110-1118.

[23] Song, S., Lan, C., Xing, J., Zeng, W., & Liu, J. (2017). An End-to-end Spatio-temporal Attention Model for Human Action Recognition from Skeleton Data. *Proceedings of the AAAI conference on artificial intelligence*, 31(1).

[24] Zhang, P., Lan, C., Xing, J., Zeng, W., Xue, J., & Zheng, N. (2017). View Adaptive Recurrent Neural Networks for High Performance Human Action Recognition from Skeleton Data. *Proceedings of the IEEE International Conference on Computer Vision*, 2117-2126.

[25] Luo, Y., Ren, J., Wang, Z., Sun, W., Pan, J., Liu, J., ... & Lin, L. (2018). Lstm Pose Machines. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5207-5215.

[26] Artacho, B., & Savakis, A. (2020). Unipose: Unified Human Pose Estimation in Single Images and Videos. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7035-7044*.

[27] Yan S, Xiong Y, Lin D. (2018). Spatial Temporal Graph Convolutional Networks for Skeleton-based Action Recognition. *Thirty-second AAAI Conference on Artificial Intelligence.*

[28] Thakkar, K., & Narayanan, P. J. (2018). Part-based Graph Convolutional Network for Action Recognition. arXiv preprint arXiv:1809.04983.

[29] Shi, L., Zhang, Y., Cheng, J., & Lu, H. (2019). Two-stream Adaptive Graph Convolutional Networks for Skeleton-based Action Recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12026-12035.

[30] Obinata, Y., & Yamamoto, T. (2021). Temporal Extension Module for Skeleton-based Action Recognition. *2020 25th International Conference on Pattern Recognition (ICPR), IEEE*, 534-540.

[31] Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., & Tian, Q. (2021). Symbiotic Graph Neural Networks for 3d Skeleton-based Human Action Recognition and Motion Prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

[32] Cui, Q., & Sun, H. (2021). Towards Accurate 3D Human Motion Prediction from Incomplete Observations. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4801-4810.

[33] Li, Z., Ye, J., Song, M., Huang, Y., & Pan, Z. (2021). Online Knowledge Distillation for Efficient Pose Estimation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11740-11750.

[34] Xian, Y., Korbar, B., Douze, M., Torresani, L., Schiele, B., & Akata, Z. (2021). Generalized Few-shot Video Classification with Video Retrieval and Feature Generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*