Research Article

Skeleton-Based Action Recognition Based on Distance Vector and Multihigh View Adaptive Networks

Min Zhang, Haijie Yang, Pengfei Li, and Ming Jiang

School of Computer Science, Hangzhou Dianzi University, Hangzhou 310000, China

Correspondence should be addressed to Haijie Yang; yhjhdu@163.com

Received 23 May 2021; Accepted 7 August 2021; Published 19 August 2021

Academic Editor: Thippa Reddy G

Copyright © 2021 Min Zhang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Skeleton-based human action recognition has attracted much attention in the field of computer vision. Most of the previous studies are based on fixed skeleton graphs so that only the local physical dependencies among joints can be captured, resulting in the omission of implicit joint correlations. In addition, under different views, the content of the same action is very different. In some views, keypoints will be blocked, which will cause recognition errors. In this paper, an action recognition method based on distance vector and multihigh view adaptive network (DV-MHNet) is proposed to address this challenging task. Among the mentioned techniques, the multihigh (MH) view adaptive networks are constructed to automatically determine the best observation view at different heights, obtain complete keypoints information of the current frame image, and enhance the robustness and generalization of the model to recognize actions at different heights. Then, the distance vector (DV) mechanism is introduced on this basis to establish the relative distance and relative orientation between different keypoints in the same frame and the same keypoints in different frames to obtain the global potential relationship of each keypoint, and finally by constructing the spatial temporal graph convolutional network to take into account the information in space and time, the characteristics of the action are learned. This paper has done the ablation study with traditional spatial temporal graph convolutional networks and with or without multihigh view adaptive networks, which reasonably proves the effectiveness of the model. The model is evaluated on two widely used action recognition benchmarks (NTU-RGB+D and PKU-MMD). Our method achieves better performance on both datasets.

1. Introduction

Human action recognition is currently one of the most important tasks in computer vision, and it is widely used in human-computer interaction, video surveillance, video understanding [1], and virtual reality [2–4]. According to the different types of data input to the network [5], human action recognition can be roughly divided into two research directions. One is the RGB-based method, which has been widely used. The other is based on the 3D skeleton method; that is, the human body is represented by joints in 3D space, which has attracted more and more attention. Past studies have shown that using 3D skeleton data to represent human movements is very effective and robust and also is conducive to computer storage and calculation [6, 7], and the skeleton data are usually obtained by the video-based pose estimation method [8].

Most of the previous methods encode the position of the joints in each frame of the video, convert them into feature vectors, and perform pattern learning [9–12]. However, these methods ignore the potential connections between joints and thus lose a lot of movement information. In order to solve this problem, a method for constructing a skeleton graph is proposed to capture the dependencies between joints, in which each joint is used as a vertex of the graph, the connection between two joints is used as an edge, and the graph convolutional network (GCN) is used to extract the features [13]. However, the method does not take the spatial and temporal information into consideration, so ST-GCN [14] is proposed to learn both spatial and temporal information. The network divides the joints into several parts to use different convolution kernels for convolution. Although this method distinguishes the joints in different regions, it
does not subdivide the relationship between the current joint and every other joints which will cause wrong recognition of some special actions. In real scenes, the height of video shooting is often different. For example, the angle of the video monitored by the camera is from top to bottom. Because of their different heights, the content presented by the same action is in variation, so training the actions data only at a single height is often not very robust.

In order to solve the above problems, this paper proposes a human action recognition method based on distance vector and multihigh view adaptive networks. First, the best observation view at different heights is automatically determined through the multihigh view adaptive networks, and the keypoints coordinates under the new observation view are formed. Then, the distance vector is calculated between each keypoint in the same frame and other keypoints and the distance vector between the same keypoints in different frames, where the distance vector includes the relative spatial distance and the relative spatial orientation. Next, spatial temporal graph convolutional networks are constructed that use joints as graph nodes and the connections between joints as edges, and finally the corresponding action category is obtained through the standard SoftMax classifier. The resulting network is illustrated in Figure 1.

The advantages of the network in this paper compared with other current networks are as follows. (1) The distance vector mechanism is introduced on the basis of the spatial temporal convolutional network, which can not only learn the information of the action in time and space but also obtain the correlation information between the keypoint and all other keypoints and realize the global association. (2) Multihigh adaptive networks are constructed to learn the keypoints information of the best observation view at different heights, under the premise of ensuring that the network is easier to learn features, and the generalization of the model is increased and the accuracy is improved. (3) The final model performs well on two popular datasets (NTU-RGB+D and PKU-MMD) and demonstrates the effectiveness of DV-MHNet.

2. Related Work

Recently, graph-based models have attracted more and more attention because of their ability to effectively represent graph data structures [15]. The current graph models mainly fall into two architectures: one framework named graph neural network (GNN), which is composed of graphs and recurrent neural networks. It is learned through multiple iterations of message passing and node states updating. It can capture the semantic relation and structural information within its neighbor nodes. In [16], the interaction between people and objects in video images is understood by using graph neural networks. In [17], GNN is used to establish dependencies between roles and predict a consistent structured output for situation recognition. The other is the graph convolutional network (GCN), which applies neural networks to the graph. Among them, the GCNs are divided into spectral GCNs and spatial GCNs. Spectral GCNs transform graph signals on graph spectral domains and then apply spectral filters on spectral domains. In [13], a semi-supervised classification based on GCN is proposed, which determines the convolutional network structure through the local first-order approximation of spectral graph convolutions. For spatial graph convolution, the convolution operation is performed through the domain information of each node to obtain the new feature vector of each node. Martin and Nikos [18] formulated a convolution-like operation on graph structure. It is a weighted average operation with the conventional two-dimensional image convolution operation. In addition to extracting spatial information, video-based action recognition must also extract temporal information. In order to capture the spatial temporal characteristics of the graph sequence, the article [19] introduces for the first time a graph convolution LSTM with a cyclic structure, which is an extension of graph convolution.

Because skeleton data are robust to illumination change and scene variation, human action recognition methods based on skeleton data have received increasing attention. These approaches can be categorized into handcrafted feature-based methods and deep learning methods. For the first method, the authors in [20] used the relative 3D rotation between each body part in each joint, and the authors in [12] used the relative 3D geometry between all pairs of body parts to represent the human body skeleton. For human action recognition based on deep learning, Du et al. [9] divided it into five parts according to the physical structure of the human body and then input them into the hierarchical recurrent neural network for action recognition. Reference [21] introduced a view adaptive model suitable for skeleton sequences, which can adjust the observation perspective itself to a suitable perspective. In the work of [6, 22, 23], it can be found that for human action recognition tasks, it is not only necessary to obtain the spatial features of the skeleton but also the temporal features, which are equally important. Based on the above work, Yan et al. [14] formulated a human action recognition method based on spatial temporal graph convolution, which can automatically learn spatial and temporal features from the data, which makes the model have strong expression and generalization capabilities. In addition, human action recognition based on local graph convolutional network [23] learns by dividing the skeleton into several parts, which better captures the structural information of the skeleton. Different from [14] and [23], [22] captures the spatial structure information by the GCN while using LSTM to model the skeleton temporal details. Although the above methods have significantly improved the performance of human action recognition, they ignore the potential relations between different joints in the same frame and the same joint in different frames, and they do not enhance the skeleton image from multiple views which result in the lower recognition accuracy.

3. Method

This paper proposes a human action recognition method based on distance vector and multihigh view adaptive network, which mainly includes three parts: multihigh view
Input video | Pose estimation | MHNet | Calculate the distance vector | ST-GCN | Softmax

---

Final class | Scores

**Figure 1:** The network architecture diagram proposed in this article including multihigh adaptive networks, calculated distance vector, and spatial temporal graph networks.

adaptive network, distance vector calculation, and spatial temporal graph convolutional network. This section first introduces the construction of multihigh view adaptive network, then explains the calculation of distance vector, and finally introduces the construction of spatial temporal graph convolutional network.

### 3.1. Multihigh View Adaptive Network

The joint content under different views is very different. For the same action, due to the change of the angle, the recognition difference is caused. For example, in the field of surveillance video, the joint information at a high view is also very different from the joint information at a horizontal view. In order to eliminate the influence caused by the difference in perspective, a variable view network architecture is constructed to automatically reobserve the skeleton from a new virtual perspective before performing action recognition. Inspired by [24], the new joints are obtained by rotating the view, but it is just to find a better perspective for the data to generate new joints coordinates and then send it to the network for classification. In this article, the coordinates of the keypoints generated under the best views of different heights are sent to the network for training, and the height is expressed by the angle with the horizontal. Figure 2 shows the best views at different heights. It can be seen that this is a bending action sequence. The first row is four frames of images extracted from the original video. The second row is the original keypoints and the corresponding human body motion modeling. The third row is the best view when the angle between the viewing angle and the horizontal is 0°. It can be found that the converted skeleton information is easier to learn, but the action has not changed. The fourth and fifth rows are the best views when the angle of view and the horizontal angle are $\theta_1$ and $\theta_2$, respectively, $\theta_2 > \theta_1$. It can be found that the best viewing angles at different heights are also different. By transforming the keypoints, the keypoints at different height views are obtained, and they are trained together to achieve the purpose of data enhancement.

Given the skeleton sequence $S$ in the three-dimensional coordinate system, the $i_{th}$ joint of the $k_{th}$ frame is expressed as $v_{k,i} = [x_{k,i}, y_{k,i}, z_{k,i}]^T$, $k \in \{1, \ldots, K\}$, $i \in \{1, \ldots, N\}$. $K$ refers to the total number of frames in a bone sequence, and $N$ refers to the total number of joints in a frame. Because the coordinates of the joints in different frames will change, in the new coordinate system, the coordinate of the keypoint is $v'_{k,i} = d_k$, where $d_k$ is the displacement of coordinate system.

A new observation view is obtained by rotating counterclockwise around the $X-axis$, $Y-axis$, and $Z-axis$; the rotation angle is $R_{k,\theta} = [\alpha_{k,\theta}, \beta_{k,\theta}, \gamma_{k,\theta}]$; all joints in the same frame share the rotation angle, so the coordinates of the joints under the new view are expressed as follows:

$$v'_{k,i,\theta} = [x'_{k,i,\theta}, y'_{k,i,\theta}, z'_{k,i,\theta}]^T = R_{k,\theta}(v_{k,i} - d_k).$$

(1)

By constructing the multihigh perspective adaptive network to automatically learn and determine the best observation views at different heights, a series of keypoint positions under new perspectives can be obtained, which can be expressed as follows:

$$V = \{v'_{k,i,\theta} | \theta \in \{0^\circ, \theta_1, \theta_2, \ldots \}\}.$$

(2)

where $\theta$ is the horizontal angle between different heights and the best view. According to the actual situation of the surveillance video, set the angle reasonably and use the DVST-GCN to recognize the skeleton information features generated by the MHNet. The MHNet is mainly composed of two LSTM network branches and a fully connected layer, one of which learns the rotation matrix of the joint where rotation matrix is $R_{k,\theta}$ and the other branch learns the displacement of joints at different heights, that is, under different $\theta$, the learned $d_k$. Finally, the three-dimensional coordinates of the joints under the new perspective are calculated by formula (1), and the schematic diagram and architecture diagram of the MHNet are shown in Figures 3 and 4.

As shown in Figure 4, the original skeleton coordinates of a certain frame of the video are input to the network and the main network is mainly composed of two LSTM branch networks. The first LSTM network is used to determine the displacement $d_k$ under the specified angle, and the other LSTM network is used to learn and determine the appropriate perspective, that is, to obtain the rotation parameters $\alpha_{k,\theta}, \beta_{k,\theta}, \gamma_{k,\theta}$. Among them, the three parameters are, respectively, expressed as the angle of rotation around the $x$, $y$, and $z$ axes. After the rotation parameters and the displacement at the specified angle are obtained, the skeleton coordinate data are rotated and transformed according to the parameters to obtain a new skeleton coordinates. Then, the new set of skeleton coordinate points and the
Figure 2: The best keypoint representation under the same action at different height angles.

Figure 3: MHNet schematic diagram. (a–c) New perspectives at different heights. Among them, (a) is the best view determined by the adaptive network, and (b) and (c) are the best view under the condition of adding angle constraints. From frame 1 to frame t the posture of the human body is constantly changing, and the MHNet can always find the best view for learning.
corresponding action labels are input into a network composed of multiple LSTMs, the skeleton data are learned from end to end, and after the features are fully connected, the output is finally obtained through the SoftMax function.

3.2. Skeleton Graph Structure. The human body is in motion for a period of time through the human pose estimation method to get the joints in each frame. Figure 5 shows the skeleton graph information of several frames. We construct a spatial temporal graph \( G = (V, E) \) on a skeleton sequence with \( N \) joints and \( K \) frames, and its node set is \( V = \{v_k|k = 1, \ldots, K, i = 1, \ldots, N\} \). The black dot in Figure 5 represents the \( i_{th} \) joint in the \( k_{th} \) frame. According to the human body structure, the joints of each frame are connected to form a spatial edge \( E_s = \{v_{ki}v_{kj}|(i, j) \in B\} \), where \( B \) is the set of all joints in the \( k_{th} \) frame, as shown in the red edge of Figure 5. The same nodes are connected in two consecutive frames into edges to form temporal edges \( E_t = \{v_{ki}v_{k(i+1)i}\} \), the green side as shown in Figure 5.

3.3. Spatial Temporal Graph Convolutional Network-Based Distance Vector. The input feature vector of the spatial temporal graph convolutional network is generally composed of the coordinate vector of the node and the estimated confidence. For the spatial temporal graph convolutional network based on the distance vector, it is necessary to consider the changes in the spatial distance and direction of all the keypoints in the same frame and all the same keypoints in different frames and describe a series of changes in the action more explicitly. The distance vector proposed in this paper is mainly composed of the distance between the keypoints and the relative direction of the keypoints. The distance between the keypoints can be expressed as follows:

\[
L(v_{mi}, v_{mj}) = \sqrt{(x_{mi} - x_{mj})^2 + (y_{mi} - y_{mj})^2 + (z_{mi} - z_{mj})^2}, \quad i, j \in (1, \ldots, N),
\]

(3)

\[
L(v_{mi}, v_{mj}) = \sqrt{(x_{mi} - x_{mj})^2 + (y_{mi} - y_{mj})^2 + (z_{mi} - z_{mj})^2}, \quad i, j \in (1, \ldots, N),
\]

(4)

where \((x, y, z)\) is the coordinate of the keypoint, and formula (3) represents the spatial distance between the \( i_{th} \) keypoint and the \( j_{th} \) keypoint in the same frame, and formula (4) represents the spatial distance between \( i_{th} \) keypoints of \( m_{th} \) frame and the \( i_{th} \) keypoint of the \( n_{th} \) frame. The relative position between the keypoints can be expressed as follows:

\[
\overrightarrow{p_{m_{i}m_{j}}} = ((x_{mi} - x_{mj}), (y_{mi} - y_{mj}), (z_{mi} - z_{mj})), \quad i, j \in (1, \ldots, N),
\]

(5)

\[
\overrightarrow{p_{n_{i}n_{j}}} = ((x_{ni} - x_{nj}), (y_{ni} - y_{nj}), (z_{ni} - z_{nj})), \quad i, j \in (1, \ldots, N),
\]

(6)

where \((x, y, z)\) is the coordinate of the keypoint and \( \overrightarrow{p} \) is a direction vector. Formula (5) represents the position of the \( j_{th} \) keypoint relative to the \( i_{th} \) keypoint in the same frame, and formula (6) represents the orientation of the \( i_{th} \) keypoint in the \( n_{th} \) frame relative to the \( i_{th} \) keypoint in the \( m_{th} \) frame. As shown in Figure 6, assuming figure (a) is the \( n_{th} \) frame image of the video and figure (b) is the \( n + 1_{th} \) frame image of the video, for the \( K_1 \) skeleton point of the \( n_{th} \) frame, the distance vector between it and other keypoints is represented by the dotted line. The arrows indicate that the dashed arrows from \( K_1 \) to \( K_2 \) indicate the distance vectors between the same keypoints in different frames. It can be seen from the figure that in the \( n + 1_{th} \) frame, the distance between the keypoint \( K_2 \) and the keypoint \( F_2 \) is closer, and the direction is slightly changed. By constructing the distance vectors between keypoints, the change information of the action can be effectively captured, and learning can be carried out through the spatial temporal graph convolutional network.

Therefore, the distance vector between keypoints, the coordinate position of the joint, and the confidence of the joint form a new input feature vector \( F_{\text{new}}(v_{mi}) \). Taking the common two-dimensional image convolution operation as a reference, the spatial temporal graph convolution operation can be written as follows:

\[
F_{\text{out}}(v_{mi}) = \sum_{v_{ni} \in B(v_{mi})} \frac{1}{Z_{mi}(v_{nj})} F_{\text{new}}(p(v_{mi}, v_{nj})) \cdot w(v_{mi}, v_{mj}),
\]

(7)

\[
F_{\text{out}}(v_{mi}) = \sum_{v_{ni} \in B(v_{mi})} \frac{1}{Z_{mi}(v_{nj})} F_{\text{new}}(p(v_{mj}, v_{mj})) \cdot w(v_{mi}, v_{nj}),
\]

where \( B(v_{mi}) \) is the domain constraint condition of the joint. For spatial graph convolution, \( B(v_{mi}) = \{v_{mj}|d(v_{mj}, v_{mi}) < D\} \), \( d(v_{mj}, v_{mi}) \) represents the shortest distance from \( v_{mj} \) to \( v_{mi} \). For temporal graph convolution, \( B(v_{mi}) = \{|v_{mj}|d(v_{mj}, v_{mi}) \leq K, |n - m| \leq \Gamma/2\} \), where \( K \) is the size of the convolution kernel and \( \Gamma \) controls the size of the convolution kernel in the temporal domain. \( p \) is the sampling function to obtain the neighbor joint centered.
on the joint, \( w \) is the weight function, and \( Z \) is the normalization item.

4. Experiment

Datasets. We evaluated our proposed method on two popular datasets (PKU-MMD and NTU-RGB+D). Among them, the PKU-MMD datasets contain more than 1,000 video clips, a total of 52 action categories, and 66 objects are completed in three camera views. The NTU-RGB+D datasets contain 56,000 video clips, covering more than 60 actions and filmed by 40 volunteers in a limited laboratory environment. The datasets provide the 3D joint positions \((x, y, z)\) in the camera coordinate system.

Experimental Settings. For better experiments, the joints are rotated by a certain angle around \( x, y, \) and \( z \). These angles are randomly generated between \(-90\) and \(90\)°. In addition, in order to simplify the experiment, we removed unnecessary keypoints. The initial learning rate of the two datasets is set to 0.005. In order to alleviate the overfitting problem, the probability of dropout is set to 0.5, the initialization parameter of the fully connected layer is set to 0, and the Adam optimizer is used to train the networks.

Evaluation Criteria. For the NTU-RGB+D datasets, the author of the datasets recommends two benchmarks: one of them is X-Sub, which includes 39889 training sets and 16390 verification sets. The training sets come from an actor’s clip and the evaluation sets clips from other performers. The other is X-View, which includes 37462 training sets and 18817 validation sets. The training sets come from the second and third views of the camera and the validation sets come from the first view of the camera. For PKU-MMD datasets, there are also two recommended evaluate protocols, i.e., X-Sub and X-view. For detection-specific hyperparameters, we basically follow the settings in [25]. In particular, we use anchor scales of \[50, 100, 200, 400\] in the temporal proposal network.

5. Results

5.1. Evaluation. Results on the NTU-RGB+D Datasets. Table 1 presents the results on the NTU-RGB+D datasets X-Sub and X-view. We have compared MH-DVNet with some advanced methods. In particular, our method is better than a series of spatial temporal graph convolutional networks, with an accuracy of 95.3% in X-Sub and 89.5% in X-view. In addition, it can be seen from the table that the accuracy of the VA-LSTM [21] method and ST-GCN [14] method is lower than that of the model proposed in this paper, which proves that this model is indeed improved on their basis.

Results on the PKU-MMD Datasets. There are few papers using this dataset, but the model has achieved good results. Table 2 shows the accuracy of the MH-DVNet on the PKU-MMD datasets. The model achieves relatively good performance with an accuracy of 92.7% in X-Sub and 94.4% in X-view.
5.2. Ablation Study. We take the ST-GCN [14] as the benchmark model and conduct an ablation study on the X-Sub and X-View verification sets of NTU-RGB+D.

**DVNet.** We evaluated the following two models together: ST-GCN (benchmark model) and DVNet; the latter is improved based on ST-GCN. The benchmark model does not use the distance vector between each keypoint and other keypoints but defines a division method to classify several joints into one category. By calculating the distance vector between each keypoint of the current frame and the distance vector of the same keypoint in different frames, the change information of the action can be effectively captured. We did a comparative test on the NTU-RGB+D verification sets, as shown in Figure 7, and we can find that our proposed method is better than the ST-GCN.

**MHNet.** In the same way, we conducted a comparative experiment on the two cases where the MHNet generates joint coordinates in multiple views and directly uses the joint coordinates in a single view. Through the MHNet, it can generate the best view of different heights, which not only ensures that the best action characteristics can be learned but also can ensure recognition accuracy under high viewing angles such as some monitoring fields.

According to the actual scene, we reasonably increased the angle of view with 10° steps each time to train the model in stages. The final model and ST-GCN and VANet were ablated experiments. The experimental results are shown in Figure 8, and it can be found that the MHNet has the best effect.

In addition, since DVNet and MHNet have achieved better results under the benchmark model, we have...
reason to merge the two into the benchmark model and conduct a comparative test on the NTU-RGB + D verification set. As shown in Figure 8, it can be found that this method is better than the benchmark model.

6. Conclusion

In this paper, an action recognition method based on MH-DVNet is proposed, which mainly succeeds in the following
three aspects. (1) The MHNNet is used to automatically determine the best observation of different heights and generate multiple joints coordinates combinations under the same action category, which improves the accuracy while also improving the specific scene robustness. (2) Using the distance vector between each keypoint to capture the change information of the action, it can learn a wealth of linkage features and improve the accuracy on the basis of the ST-GCN. (3) We have done a lot of ablation experiments and reasonably proved the effectiveness of MH-DVNet.

7. Further Study

The method proposed in this paper has good robustness to the action information from different perspectives, so the complexity of the model is relatively large. In the future, we will try to reduce the complexity of the model without reducing the accuracy. In addition, this model can further improve the accuracy; a one-hot encoding technique [35] is used to convert the categorical data values to binary form. This is followed by the implementation of a crow search algorithm (CSA) for selecting optimal hyperparameters for training of dataset using the convolution neural networks. This method can be used to transform the action classification information into binary information and train a better model. Furthermore, a novel multidirectional long short-term memory (MLSTM) technique [36] is being proposed to predict the stability of the smart grid network. It has higher accuracy than the traditional LSTM, so it can be used to determine the best view of action, so as to improve the accuracy.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by Zhejiang Provincial Technical Plan Project (No. 2020C03105).

References

[1] R. Poppe, “A survey on vision-based human action recognition,” Image and Vision Computing, vol. 28, no. 6, pp. 976–990, 2010.
[2] Z. Duric, W. D. Gray, R. Heishman et al., “Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction,” Proceedings of the IEEE, vol. 90, no. 7, pp. 1272–1289, 2002.
[3] U. Gaur, Y. Zhu, B. Song, and A. Roy-Chowdhury, “A string of feature graphs model for recognition of complex activities in natural videos,” in Proceedings of the International Conference on Computer Vision (ICCV), pp. 2595–2602, Barcelona, Spain, November 2011.
[4] M. R. Sudha, K. Sriragah, S. S. Abisheck, S. G. Jacob, and S. Manisha, “Approaches and applications of virtual reality and gesture recognition: a review,” International Journal of Ambient Computing and Intelligence, vol. 8, pp. 1–18, 2017.
[5] D. Weinland, R. Ronfard, and E. Boyer, “A survey of vision-based methods for action representation, segmentation and recognition,” Computer Vision and Image Understanding, vol. 115, no. 2, pp. 224–241, 2011.
[6] C. Li, Q. Zhong, D. Xie, and S. Pu, “Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation,” in Proceedings of the 27th International Joint Conference on Artificial Intelligence IJCAI’18, pp. 786–792, Stockholm Sweden, July 2018.
[7] Y. Yan, J. Xu, B. Ni, W. Zhang, and X. Yang, “Skeleton-aided articulated motion generation,” in Proceedings of the ACM International Conference on Multimedia (ACMMM), pp. 199–207, Mountain View, CA, USA, October 2017.
[8] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, “Realtime multiperson 2d pose estimation using part affinity fields,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7291–7299, July 2017.
[9] Y. Du, W. Wang, and L. Wang, “Hierarchical recurrent neural network for skeleton based action recognition,” in Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1110–1118, Boston, MA, USA, June 2015.
[10] B. Fernando, E. Gavves, J. M. Oramas, A. Ghodrati, and T. Tuytelaars, “Modeling video evolution for action recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5378–5387, Boston, MA, USA, June 2015.
[11] M. Liu, H. Liu, and C. Chen, “Enhanced skeleton visualization for view invariant human action recognition,” Pattern Recognition, vol. 68, pp. 346–362, 2017.
[12] R. Vemulapalli, F. Arrate, and R. Chellappa, “Human action recognition by representing 3d skeletons as points in a lie group,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 588–595, Columbus, OH, USA, June 2014.
[13] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in Proceedings of the 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 2017.
[14] S. Yan, Y. Xiong, and D. Lin, “Spatial temporal graph convolutional networks for skeleton-based action recognition,” in Proceedings of the 32nd AAAI Conference on Artificial Intelligence, pp. 7444–7452, Palo Alto, CA, USA, February 2018.
[15] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, “How powerful are graph neural networks?,” 2018, http://arxiv.org/abs/1810.00826.
[16] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu, “Learning human-object interactions by graph parsing neural networks,” in Proceedings of the 15th European Conference on Computer Vision, Munich, Germany, September 2018.
[17] R. Li, M. Tapaswi, R. Liao, J. Jia, R. Urtasun, and S. Fidler, “Situation recognition with graph neural networks,” in Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, October 2017.
[18] S. Martin and N. Komodakis, “Dynamic edgeconditioned filters in convolutional neural networks on graphs,” in Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, July 2017.
[19] Y. Seo, M. Defferrard, P. Vandergheynst, and X. Bresson, “Structured sequence modeling with graph convolutional recurrent networks,” 2016, http://arxiv.org/abs/1612.07659.

[20] R. Vemulapalli and R. Chellappa, “Rolling rotations for recognizing human actions from 3d skeletal data,” in Proceedings of the 2016 Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, June 2016.

[21] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, “View adaptive recurrent neural networks for high performance human action recognition from skeleton data,” in Proceedings of the 2017 International Conference on Computer Vision, Venice, Italy, October 2017.

[22] C. Si, Ya Jing, W. Wang, L. Wang, and T. Tan, “Skeleton-based action recognition with spatial reasoning and temporal stack learning,” in Proceedings of the 15th European Conference on Computer Vision, Munich, Germany, September 2018.

[23] K. Thakkar and P. J. Narayanan, “Part-based graph convolutional network for action recognition,” in 2018 British Machine Vision Conference, Newcastle, UK, September 2018.

[24] P. Zhang and C. Lan, “View adaptive neural networks for high performance skeleton-based human action recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, 2019.

[25] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: towards real-time object detection with region proposal networks,” News in Physiological Sciences, pp. 91–99, 2015.

[26] 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 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1010–1019, Las Vegas, NV, USA, June 2016.

[27] J. Liu, A. Shahroudy, Xu Dong, and G. Wang, “Spatio-temporal lstm with trust gates for 3d human action recognition,” in Proceedings of the 14th European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, October 2016.

[28] H. Wang and L. Wang, “Modeling temporal dynamics and spatial configurations of actions using twostream recurrent neural networks,” in Proceedings of the 2017 Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, July 2017.

[29] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, “An end-to-end spatio-temporal attention model for human action recognition from skeleton data,” in Proceedings of the Thirty-First Conference on Artificial Intelligence, San Francisco, CA, USA, February 2017.

[30] I. Lee, D. Kim, S. Kang, and S. Lee, “Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks,” in Proceedings of the 2017 International Conference on Computer Vision, Venice, Italy, October 2017.

[31] C. Si and W. Chen, “An attention enhanced graph convolutional LSTM network for skeleton-based action recognition,” in Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, June 2019.

[32] Y. Li, C. Lan, J. Xing, W. Zeng, C. Yuan, and J. Liu, “Online human action detection using joint classification regression recurrent neural networks,” in Proceedings of the 14th European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, October 2016.

[33] Bo Li, H. Chen, Y. Chen, Y. Dai, and M. He, “Skeleton boxes: solving skeleton based action detection with a single deep convolutional neural network,” in Proceedings of the 2017 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), pp. 613–616, Hong Kong, China, July 2017.

[34] C. Li, Q. Zhong, Di Xie, and S. Pu, “Skeleton-based action recognition with convolutional neural networks,” in Proceedings of the 2017 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), pp. 597–600, Hong Kong, China, July 2017.

[35] T. R. Gadekallu, M. Alazab, R. Kaluri et al., “Hand gesture classification using a novel CNN-crow search algorithm,” Complex and Intelligent Systems, no. 6, 2021.

[36] M. Alazab, S. Khan, S. Krishnan et al., “A multidirectional LSTM model for predicting the stability of a smart grid,” IEEE Access, no. 99, p. 1, 2020.