Proceedings of CECNet 2021
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA210435

Algebra Based Human Skeleton Sequence Reduction and Action Recognition

Shibin XUAN [a,b,1] , Kuan WANG [a] , Lixia LIU [a] , Chang LIU [a] and Jiaxiang LI [a]
[a] *School of Artificial Intelligence, Guangxi University for Nationalities, China*
[b] *Guangxi Key Laboratory of Hybrid Computation and IC Design Analysis, China*

Abstract. Skeleton-based human action recognition is a research hotspot in recent years, but most of the research focuses on the spatio-temporal feature extraction by convolutional neural network. In order to improve the correct recognition rate of these models, this paper proposes three strategies: using algebraic method to reduce redundant video frames, adding auxiliary edges into the joint adjacency graph to improve the skeleton graph structure, and adding some virtual classes to disperse the error recognition rate. Experimental results on NTU-RGB-D60, NTU-RGB-D120 and Kinetics Skeleton 400 databases show that the proposed strategy can effectively improve the accuracy of the original algorithm.

Keywords. Action recognition, redundant frames reduction, joint adjacency graph, virtual classes

1. Introduction

Human action recognition has always been one of the research hotspots in computer vision. In recent years, human action recognition has attracted a large number of researchers' interest because of its wide application fields, such as intelligent video surveillance, advanced human-computer interaction, virtual reality, taking care for the old man and the children, computer-assisted teaching and intelligent assisted driving, etc. In the development of human behavior recognition technology, researchers have put forward many methods, among which the typical representatives are: the action recognition method based on tracking technology [1,2], the action recognition method based on human column structure [3-5], action recognition method based on spatio-temporal descriptor [6-8], Recognition method based on image depth information [9,10], the action recognition method based on "word bags" [11,12] and the action recognition method based on human skeleton [13,14,15]. Because the human skeleton data has the advantages of not being affected by human body shape and background, and the decrease of the price of depth sensors used to acquire human skeleton data and the emergence of skeleton data generation algorithms, the research of human action recognition algorithm based on skeleton data become a research hotspot. The skeleton-based action recognition deep learning method has attracted a large number of researchers' attention.

At present, deep learning models of skeleton-based human action recognition mainly include: Recurrent Neural Networks (RNN)[16,17,18,19,20], Convolutional

[1] Corresponding Author, Shibin Xuan E-mail: sbinx@qq.com.

Neural Networks(CNN)[21,22,23], and Graph Convolutional Networks(GCN) [18, 24, 25,26], Semantics-Guided Neural Networks[27].

Although these methods have achieved great success, they mostly learn the relationship between nodes in a frame through convolutional neural network and the change relationship between frames on the time axis through convolutional neural network or recursive neural network. Due to the diverse characteristics of human movement, even if the same movement is made by different people in different time periods, there are also great differences. At present, most methods are to expand and rotate the position of the key node, but the speed difference of action is only solved by multi-size window, which is obviously only a trade-off. In fact, it can be seen from the case analysis that human movement is completed through the activities of limbs and trunk, and each movement of limbs has its starting point and ending point. The number of video frames between the starting point and ending point of limbs movement in the slow video segment are more than those in the fast one, which increases the uncertainty of the convolution calculation on the time axis. For this purpose, in this paper, we propose a method based on algebraic operation to reduce the number of frames between the starting frame and the ending frame, and try to reduce the inconsistency caused by the different speed of action. At the same time, in order to improve the relative relationship between nodes, some connections between nodes that are not physically adjacent are added to the skeleton graph. In addition, the number of target classes is increased to disperse the error rate. Since the proposed measures do not involve any network structure, they can be applied to any existing skeleton-based human action recognition models.

2. Introduction of related basic work

The earliest and most successful application of graph convolutional neural network in skeleton-based human behavior recognition is ST-GCN [15], where the skeleton data is represented as a spatio-temporal graph $G = (V, E)$, in which the vertices in the graph is correspond to the human joints, and the edge represents the human bones and the line between same joints in the consecutive frame. A video segment can be represented as $X \in R^{N\times T\times C}$, where N represents the number of joints in an frame image, T represents the number of frames in a video, and C represents the number of channels. A relations graph between joints in a single frame can be represented by an adjacency matrix $A \in \{0,1\}^{N\times N}$, I denotes the identity matrix. $A + I$ can be divided into three parts $A_j (j = 1,2,3)$ according to the three cases {root, centripetal, earth}, and satisfies $A + I = \sum_{j=1}^{3} A_j$. In the case of single frame, spatial convolution calculation can be expressed as:

$$F_{out} = \sum_{j=1}^{3} \Lambda_j^{-\frac{1}{2}} A_j \Lambda_j^{-\frac{1}{2}} F_{in} W_j \qquad (1)$$

Where $\Lambda_j^{ii} = \sum_k (A_j^{ik}) + \alpha$, α is a regulating constant, in generally set as $\alpha = 0.001$.

Ke Cheng et al. [28] introduced the shift-CNNS idea into ST-GCN and proposed shift-GCN, which greatly reduced the model complexity. Shift-GCN includes Shift space graph convolution and Shift time graph convolution. Shift space graph convolution is divided into local Shift graph convolution and non-local Shift graph convolution. The neighborhood of node v is denoted as $B_v = \{B_v^1, B_v^2, \cdots, B_v^n\}$, and the channel at node v is divided into $n + 1$ parts. Among them, the first part retains the

characteristics of node v, and the other n parts are gained from shifting $B_v^1, B_v^2, \cdots, B_v^n$ respectively. If $F \in R^{N\times C}$ is supposed to be a human skeleton feature in a video frame, the corresponding shift features $\tilde{F} \in R^{N\times C}$ can be calculated according to the following formula:

$$\tilde{F} = F_{(v,:c)} \parallel F_{(B_v^1,c:2c)} \parallel F_{(B_v^2,2c:3c)} \parallel \cdots \parallel F_{(B_v^n,nc:)} \quad (2)$$

Where $c = \left\lfloor \frac{C}{n+1} \right\rfloor$, $\parallel$ represents channel-wisely cascade, and a learnable mask is used to realize non-local Shift graph convolution.

Ziyu Liu[29] et al. proposed loose graph convolution network and consistent spatio-temporal graph convolution operator to solve the unbiased long-domain connection relation under multi-scale operator and barrier-free capture of complex spatio-temporal dependence from spatio-temporal information flow. For this purpose, First of all, the k adjacency matrix $\tilde{A}_{(k)}$ can be defined as follows:

$$\left[\tilde{A}_{(k)}\right]_{i,j} = \begin{cases} 1 & \text{if } d(v_i, v_j) = k, \\ 1 & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases} . \quad (3)$$

Then the output of the spatial convolution model can be expressed as:

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

Let the size of the sliding window along the time axis be τ and the window expansion rate be d,

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

Similarly, we can define joint features $X_{(\tau,d)} \in R^{T\times \tau N\times C}$.

The final MS-G3D model compute can be expressed as:

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

3. Skeleton data reduction based on algebraic computation and category optimization

From the last section 2, their work is more about network structure improvement, but network structure improvement has certain limits. We look forward to improving the correct recognition rate without changing the network structure. In this part, we will introduce in detail how to use algebraic method to reduce the video frames on the time axis, analyze the changing characteristics of the relationship between joints in human activities and the theoretical basis of category optimization.

3.1. Framework frame data reduction based on algebraic calculation

Although the human skeleton data has greatly simplified the video data, because the human body motion curve described by the skeleton is still a high-dimensional curve, the motion difference makes the motion curve of the same action also have a great difference. In addition, the start and end frames of the obtained skeleton data are also inconsistent, and the number of frames expressing a complete action is also

inconsistent. All these make the activity recognition based on skeleton data have great uncertainty.

Human action is formed by the movement of human limbs and head, and the movement of limbs is the key factor of action recognition. No matter how complex the movements of the human body are, the movements of the limbs are made up of a combination of fragments of movement in a particular direction, and the characteristics of movement are mainly determined by the starting and ending points of this movement in a particular direction. The starting and ending video frames play a key role in motion recognition, and the video frames determining the activity category are referred to as key frames.

As the skeleton data of video sequence is a high-dimensional data stream, in general, it is difficult to directly calculate the corresponding key frame for such high-dimensional data. However, if we can establish a mapping from such high-dimensional data to one-dimensional data, and this mapping can retain the trend characteristics of the original human movement, we can easily obtain the corresponding video key frames from this one-dimensional data. After many experiments we found that there is a characteristic function that has this function. Here, we present a mapping function that can be used for fast and efficient key frame extraction.

Assuming that the number of joints in the skeleton is n, x_{ij} represents the jth joint node of ith frame, and the internal structure matrix of kth frame is expressed as follows:

$$R_k = (r_{st}^k)_{n\times n} \qquad (7)$$

Where $r_{st}^k = dist(x_{ks}, x_{kt}), s, t = 1, \cdots, n$, the function $dist(\cdot,\cdot)$ represents the distance between two points. Then the mapping function of $\mathbb{R}^{n\times n} \to \mathbb{R}$ is defined as:

$$y_k = smooth(std(eig(R_k))). \qquad (8)$$

Where the function $eig(\cdot)$ represents the eigenvalue vector of the matrix, $std(\cdot)$ denotes the standard variance of the vector, and $smooth(\cdot)$ refers to as the data smoothing process.

Let $S_{max} = maxid([y_k]), S_{min} = minid([y_k]), S_{ext} = uion(S_{max}, S_{min})$, where $maxid(\ )$ represents the indies of the maximum value, $minid(\ )$ denotes the indies of the minimum value, and $uion(\ )$ is the set union operation, so that set S_{ext} includes the indies set of the key frames in the corresponding video.

For example, we select a video in MSRAction 3D database and solve key frames according to the above calculation formula. The mapping function curve of the video segment in Figure 1 is shown in Figure 2, here Minimum frame numbers: 4, 16, 26, 37, Maximum frame numbers: 1, 12, 21, 27.

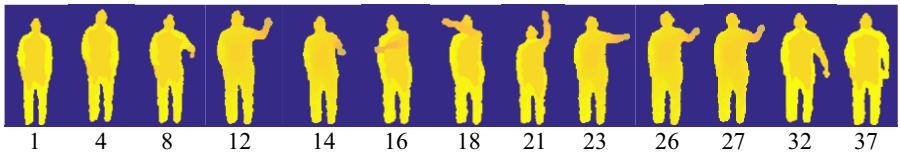

Figure 1. The depth information image of a video with extreme frames in MSRAction 3D

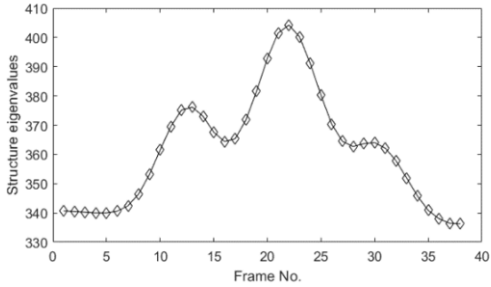

Figure 2 Mapping function of video segment in Fig. 1

It can be seen from Fig.1 that the mapping function defined in this way from skeleton data to motion trend basically meets the segmentation requirements of action units.

In order to further illustrate the rationality of the mapping function established above, we randomly generate a symmetric matrix with a size of 20×20, multiply the values of the kth row and the kth column by a periodic change function such as $\sin(\pi x)$, and let $x = 0.1,0.2,0.3,\cdots,6$, then the mapping function curve is obtained as shown in Fig.3. Because when a joint of human body is moving, the distance between this node and other joints will change, the row and column values corresponding to this joint in the corresponding distance matrix will change accordingly. Since the human

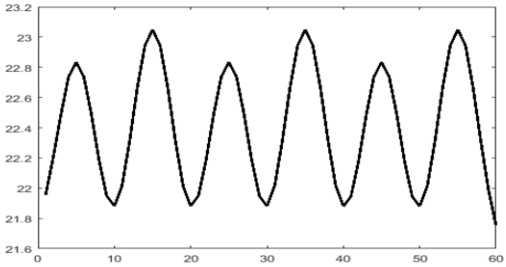

Figure.3 The mapping function curve of the random symmetric matrix is obtained by multiplying the 5th row and 5th column by the periodic function sin(πx).

body movement is in a specific direction at a certain time, it represents the increase or decrease of values in the same row and column in the structural matrix. Fig.3 shows that the structure mapping function defined can better reflect human body changes.

Advantages of the algorithm: fast calculation speed. Generally, the number of joints in a frame is 15, 20 or 25. The corresponding distance matrix does not exceed 25×25 dimensions. For such a low dimension, it takes very little time for the computer to calculate its eigenvalue and the corresponding variance, filtering and extreme value. We summarize the above calculation process into the following algorithm.

Algorithm 1 Video frame reduction

Input: original video sequence X, smoothing filter smooth(), extremum extraction function maxid(), minid().

Output: reduced video sequence Y.

Step 1: Compute structure matrix R_k of kth frame by formula 7.

Step2: Compute feature value y_k of structure matrix R_k of kth frame by formula 8.

Step3: Compute extremum frame number set S_{ext} of X by functions $maxid()$,$minid()$, $uion()$.
Step4: Output the frames set of X with frame number in S_{ext}.

3.2. Skeleton graph

In most models, the vertices in the skeleton structure graph are taken as each joint and the edges as the physiological bone. In some new models, some new edges are added on the natural skeleton graph to express the importance of the position relations between these joints, so as to make up for the fact that the convolution operation of the natural skeleton graph cannot take into account these relations. For example, Yansong Tang [24] et al., in order to show the importance of the relative position relation between the joints of limbs for activity description, added the connection between the corresponding joints of limbs on the basis of the original natural connection, as shown in Fig.4.

Considering the importance of limbs in motion, in addition to the natural connection between nodes, this paper adds the connection edge between some limbs nodes and the central position of the human body to enhance the role of limbs in the representation of human posture. See Fig.5

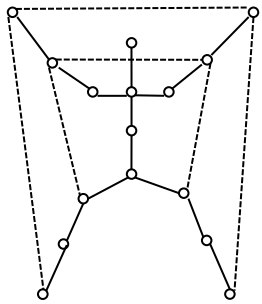

Figure 4. Solid lines represent bones and dotted lines represent new added edges.

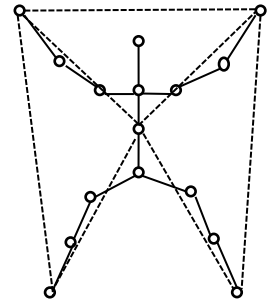

Figure 5. Solid lines represent bones and dotted lines represent new added edges.

3.3. Dispersing error rate

At present, almost all algorithm models adopt one-hot mode when they express the sample target category, that is, the target category is set as a vector y with length N of the total number of categories, where the ith element of vector y representing class i is 1 and other elements are 0. It is found in the experiment that if the length of vector y is doubled, that is, $2N$, and the ith element of vector y representing class i is 1, and all other elements are 0, the model accuracy will be improved. For the test sample x with class i, whether the system model can correctly identify sample x can be judged according to whether its output value on sample x is i. Therefore, in the case of expanding the number of categories, even if the output of the system model on the input sample x exceeds the initial sample category range, it only indicates that the system cannot correctly identify sample x, but the system can still operate normally. The ith element of the category vector y_x of the sample x of class i is 1 and the rest

elements are 0. According to the output structure of the model, when training the model with the training sample x of class i, it is expected that the model output is as close as possible to the class vector y_x of sample x, that is, the ith element of the output vector $\hat{y}_x$ of the model should be as close to 1 as possible, and other elements should be as close to 0 as possible. For any test sample x', when it is fed into the trained model, it is assumed that the probability of its output correct category is p and the error probability is q. When $p \gg q$, the model outputs almost the correct category, but when p approaches q or is less than or equal to q, the model outputs the wrong category more likely. The model output should be one of the other $N-1$ categories of the incorrect category, denoted j. The probability of the model output class J is $q/(N-1)$ under the assumption of equal probability. If the length of the class vector y becomes $2N$ without changing the total number of classes, the model outputs class j with a probability of $q/(2N-1)$. Thus, under the assumption of equal probability, appropriately extending the length of category vector may reduce the probability of outputting a certain error category. However, experiments show that the length of category vector should not be increased too much, which will reduce the accuracy of model output. The reason may be that adding too many categories reduces the probability of output correct categories while reducing the error categories of the model. Therefore, in this paper, all experiments increase the length of category vector by 1 or 0.5 times.

4. Analysis of experimental results

To demonstrate the effectiveness of the proposed approach, three databases commonly used to evaluate skeleton-based human recognition models, NTU-RGB-D60, NTU-RGB-D120, and Kinetics Skeleton 400, were selected as benchmark databases. All video segments of the database are processed by the method of eliminating redundant frames introduced in section 3.1, and the empty frames at the tail are filled with zero. Four models ST-GCN[15], 2s-AGCN[26], SGN[27], Shift-GCN[28] and MS-G3D[29] were selected as the benchmark models.

4.1. Experiments on NTU-RGB-D60 database

NTU-RGB-D60[30] contains 56,578 skeleton sequences and 60 action categories, all of which were completed by 40 different people and captured by 3 cameras from different perspectives. Each frame contains the three-dimensional coordinates of 25 body nodes. In cross-subject (X-sub), the training set contained 40,091 samples and the test set contained 16,487 test instances. In cross-view (X-view), the test set consisted of all 18,932 samples collected by camera 1, and the remaining 37,646 samples were used for training. All experimental results are listed in Table 1.

Table 1 Comparison of Top1 results on NTU-RGB-D60 database

The algorithm name	Xsub		Xview	
	Original	+our	Original	+our
ST-GCN	81.5	85.3	88.3	91.8
2sAGCN	88.5	89.3	95.1	95.7
SGN	89.0	89.5	94.5	95.1
Shift-GCN	90.7	91.0	96.5	96.5
MS-G3D	91.5	91.7	96.2	96.3

It can be seen from Table 1 that the proposed method improves the recognition results on the basis of the original algorithm, and greatly improves the recognition rate

of the basic GCN algorithm ST-GCN, and the correct recognition rate of Xsub and Xview are increased by 4.66% and 3.96% respectively. However, the improvement of correct recognition rate of MS-G3D algorithm is limited. The possible reason is that there is a limit on the correct recognition rate of GCN-based algorithm on this database, and there is not much range for improvement of correct recognition rate of algorithm.

4.2. Experiments on NTU-RGB-D120 database

NTU RGB+D 120[31] added an additional 57,367 skeleton sequences on the basis of NTU RGB+D 60, with a total sample number of 113,945, including more than 120 categories. In CrossSetup (X-set), the training Set contains 54,468 samples collected by half of the cameras, with the remaining 59,477 samples used for testing. In cross-subject, the training set consisted of 63,026 samples from 53 subjects, and the remaining 50,919 samples were used for testing. All experimental results are listed in Table 2.

Table 2 Comparison of Top1 results on NTU-RGB-D120 database

The algorithm name	Xsub		Xset	
	Original	+our	Original	+our
2sAGCN	82.9	84.5	84.9	86.3
SGN	79.2	81.0	81.5	83.2
Shift-GCN	85.9	86.7	86.2	87.4
MS-G3D	86.9	87.3	87.6	88.1

NTU-RGB-D120 has one times more activity categories than NTU-RGB-D60, and the correct recognition rate of each algorithm is obviously lower than that of NTU-RGB-D60. However, the correct recognition rate of these algorithm models is improved with the measures proposed by us. Table 2 shows that the correct recognition rate of 2sAGCN increases by 1.92% and 1.65% in Xsub and Xview respectively after adding our proposed strategy.

4.3. Experiments on Kinetics Skeleton 400 database

The Kinetics Skeleton 400[32] dataset contained 240,436 training samples and 19,796 test Skeleton sequences, and over 400 classes. The experimental results on this database are shown in Table 3.

Table 3 Comparison of Top1 results on Kinetics Skeleton 400 database

The algorithm name	Top-1(%)		Top-5(%)	
	Original	+our	Original	+our
ST-GCN	30.7	33.5	52.8	56.4
2sAGCN	36.1	37.0	58.7	59.3
MS-G3D	38.0	38.5	60.9	61.2

The Kinetics Skeleton 400 DATASET contains much higher activity categories than NTU-RGB-D. Table 3 shows that the performance of each algorithm is very sensitive to activity categories, and the correct recognition rate reduces significantly. However, the proposed strategy can still effectively improve the performance of each algorithm. Table 3 shows that the correct recognition rate of ST-GCN improved by 9.1% and 6.8% in Top-1 and Top-5 respectively after adding our proposed strategy.

5. Conclusion

In this paper, we propose a new method of video frame reduction, which transforms a high-dimensional video frame sequence into a one-dimensional time series by algebraic method. Experiments show that the one-dimensional sequence can effectively represent the movement trend of limb movement, and the extreme point of the sequence corresponds to the time when the movement trend changes, while the non-extreme point of the sequence has little significance for the identification of the whole movement. In this way, we can select the video frame corresponding to the extreme point of the one-dimensional sequence to form a new sequence, so as to achieve the purpose of reducing the redundant video sequence signal. At the same time, we added lines between some non-physiologically connected joints in the human skeleton diagram to enhance the ability of representing actions by the relative position relationship between key joints. Finally, the false recognition rate is reduced by adding virtual categories. Experimental results show that our strategy can effectively improve the correct recognition rate of the original algorithm model.

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

References

[1] WANG H,KLASER A,SCHMID C,et al.Action recognition by dense trajectories[C]//Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, Jun 20-25, 2011. IEEE, 2011: 3169-3176.
[2] WANG H, SCHMID C. Action recognition with improved trajectories[C]//International Conference on ComputerVision, Sydney,NSW,Australia, Dec 1-8, 2013. IEEE, 2014: 3551-3558.
[3] D. Marr, H.K. Nishihara. Representation and recognition of the spatial organization of three-dimensional shapes. Proceedings of the Royal Society of London. Series B. Biological Sciences, 1978, 200 (1140): 269-294.
[4] D. Hogg. Model-based vision: a program to see a walking person. Image and vision computing, 1983, 1 (1): 5-20.
[5] K. Rohr. Towards model-based recognition of human movements in image sequences. CVGIP-Image Understanding, 1994, 59 (1): 94-115.
[6] LAPTEV I.On space-time interest points[J]. International JournaI of Computer Vision, 2005, 64(2-3): 107-123.
[7] WANG H, ULLAH M M, KLASER A, et al. Evaluation of local spatio-temporal features for action recognition[C]//CAVALLARO A, British Machine Vision Conference, 2009.BMVA,2009: 124.1 - 124.11.
[8] WILLEMS G, TUYTELAARS T, GOOL L V. An efficient dense and scale-invariant spatio-temporal interest point detector[C]//FORSYTH D, 2008 European Conference on Computer Vision, Springer, Berlin, Heidelberg. 2008,5303: 650-663.
[9] YANG X D, TIAN Y L. Effective 3D action recognition using eigenjoints[J]. Journal of Visual Communication and Image Representation, 2014, 25(1): 2-11.
[10] ZHANG H X, YE Y S, CAI X Z, et al. Efficient algorithm of action recognition based on joint points[J]. Computer Engineering and Design, 2020, 41(11) : 3168-3174
[11] Wangqing Li, Zhenyu Zhang, Zicheng Liu. Li, W., Zhang, Z. & Liu, Z. (2010). Action recognition based on a bag of 3D points. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010 (pp. 9-14). Piscataway, New Jersey, USA: IEEE.
[12] Muhammet Fatih Aslan, Akif Durdu,Kadir Sabanci. Human action recognition with bag of visual words using different machine learning methods and hyperparameter optimization. Neural Computing and Applications, 32,8585–8597(2020).

[13] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In CVPR, 2015.
[14] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. A new representation of skeleton sequences for 3d action recognition. In CVPR, 2017.
[15] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In AAAI, 2018.
[16] Wentao Zhu, Cuiling Lan, Junliang Xing, Wenjun Zeng, Yanghao Li, Li Shen, and Xiaohui Xie. Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In AAAI, 2016.
[17] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In ICCV, 2017.
[18] Chenyang Si, Ya Jing, Wei Wang, Liang Wang, and Tieniu Tan. Skeleton-based action recognition with spatial reasoning and temporal stack learning. In ECCV, 2018.
[19] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In CVPR, 2016
[20] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In ECCV, 2016.
[21] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View adaptive neural networks for high performance skeleton-based human action recognition. TPAMI, 2019.
[22] Junwu Weng, Mengyuan Liu, Xudong Jiang, and Junsong Yuan. Deformable pose traversal convolution for 3d action and gesture recognition. In ECCV, 2018.
[23] Congqi Cao, Cuiling Lan, Yifan Zhang, Wenjun Zeng, Hanqing Lu, and Yanning Zhang. Skeleton-based action recognition with gated convolutional neural networks. TCSVT, 29(11):3247–3257, 2019.
[24] Yansong Tang, Yi Tian, Jiwen Lu, Peiyang Li, and Jie Zhou. Deep progressive reinforcement learning for skeleton-based action recognition. In CVPR, 2018.
[25] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. arXiv, 2016.
[26] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Twostream adaptive graph convolutional networks for skeletonbased action recognition. In CVPR, 2019 In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 12026–12035, 2019.
[27] Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, Nanning Zheng. Semantics-Guided Neural Networks for Efficient Skeleton-Based Human Action Recognition. In CVPR2020.
[28] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, Hanqing Lu. Skeleton-Based Action Recognition with Shift Graph Convolutional Network.In CVPR2020.
[29] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, Wanli Ouyang.The Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition, In CVPR2020.
[30] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang 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, pages 1010–1019, 2016.
[31] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. NTU RGB+D 120: A large-scale benchmark for 3d human activity understanding. CoRR, abs/1905.04757, 2019.
[32] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017.