Electronic Research Archive

http://www.aimspress.com/journal/ERA

ERA, 30(11): 4157–4177.
DOI: 10.3934/era.2022210
Received: 29 July 2022
Revised: 29 August 2022
Accepted: 01 September 2022
Published: 21 September 2022

Research article

Video behavior recognition based on actional-structural graph convolution and temporal extension module

Hui Xu[1], Jun Kong[1,2], Mengyao Liang[1], Hui Sun[3,*] and Miao Qi[1,2,*]

[1] College of Information Science and Technology, Northeast Normal University, Changchun 130117, China
[2] Key Laboratory of Applied Statistics of MOE, Northeast Normal University, Changchun 130024, China
[3] Institute for Intelligent Elderly Care, Changchun Humanities and Sciences College, Changchun 130117, China

* **Correspondence:** Email: sunh333@nenu.edu.cn, qim801@nenu.edu.cn; Tel: +13756563665, +13843062946.

Abstract: Human behavior recognition has always been a hot spot for research in computer vision. In this paper, we propose a novel video behavior recognition method based on Actional-Structural Graph Convolution and a Temporal Extension Module under the framework of a Spatio-Temporal Graph Convolution Neural Network, which can optimize the spatial and temporal features simultaneously. The basic network framework of our method consists of three parts: spatial graph convolution module, temporal extension module and attention mechanism module. In the spatial dimension, the action graph convolution is utilized to obtain abundant spatial features by capturing the correlations of distant joint features, and the structural graph convolution expands the existing skeleton graph to acquire the spatial features of adjacent joints. In the time dimension, the sampling range of the temporal graph is expanded for extracting the same and adjacent joints of adjacent frames. Furthermore, attention mechanisms are introduced to improve the performance of our method. In order to verify the effectiveness and accuracy of our method, a large number of experiments were carried out on two standard behavior recognition datasets: NTU-RGB+D and Kinetics. Comparative experiment results show that our proposed method can achieve better performance.

Keywords: skeleton-based behavior recognition; graph convolution network; temporal extension module; attention mechanism; spatio-temporal features

1. Introduction

The volume of video data has significantly increased in recent years, which provides great amounts of data for video behavior recognition [1]. However, the complexity of human posture, variations in view and background interference affect the recognition accuracy. Collection of skeletal data has become simpler due to continued improvement of depth cameras and human posture prediction algorithms. In addition, skeleton information has certain robustness to background, illumination and occlusion, and thus it is widely used in human behavior recognition studies.

Currently, there are two general categories of behavior recognition methods: traditional manual design methods [2] and behavior recognition methods based on deep learning [3]. The principle of the traditional manual design method is to manually extract the spatial and temporal features that represent human behavior [4]. Traditional methods mainly utilize machine learning methods such as the Support Vector Machine (SVM) and Probability Graph model for behavior recognition. Zhang et al. [5] developed a motion-based model known as Motion Context using image representation techniques, and they established a human behavior template for behavior matching. Niebles et al. [6] proposed an unsupervised learning method for human action categories by extracting space-time interest points of behavioral changes. Wang et al. [7] designed a video model based on dense trajectories and motion boundary descriptors to capture the local motion information of the video. The above methods are robust to occlusion and illumination, but they cannot handle view variations with high computational complexity and low speed.

Early behavior recognition methods based on the skeleton used the traditional methods to manually extract skeletal data to simulate the behavior of the human body. Vemulapalli et al. [8] developed a skeletal model that explicitly simulated the 3D geometric relationships between various body parts using rotations and translations in 3D space. Hussein et al. [9] used a covariance matrix for skeleton joint locations over time as a discriminative descriptor for a sequence to recognize human action. Ofli et al. [10] represented human actions by automatically selecting a few skeletal joints that were deemed to be the most informative based on highly interpretable measures such as the mean or variance of joint angle trajectories. Xia et al. [11] utilized histograms of 3D joint locations as the representations of postures for human action recognition. The methods based on artificial skeleton extraction no longer meet the requirements of high precision. The application of deep learning models can automatically extract features and avoid the complexity and differences of manual features. Thus, behavior recognition methods based on deep learning are an important development trend.

At present, the skeleton behavior recognition methods based on deep learning generally focus on constructing a skeleton graph with joints as vertices and skeletons as edges. A Convolution Neural Network (CNN) [12–15] is then utilized to extract the spatial and temporal features of behavior. These methods can extract the spatial features of adjacent joints and the temporal features of the same joints in adjacent frames. However, the conventional methods mainly involve optimization of spatial maps and disregard the optimization of temporal maps. In the time dimension, these methods only obtain the correlation between the same joints in adjacent frames without considering the relationships between adjacent joints in adjacent frames.

Aiming at the above problems, we propose a novel video behavior recognition method based on Actional-Structural Graph Convolution and a Temporal Extension Module, which can simultaneously optimize the spatial and temporal features. Inspired by the Spatio-Temporal Graph Convolution Neural Network (STGCN), the Actional-Structural Graph Convolution [16] is exploited to extract relevant features of distant joints in the spatial dimension, to achieve the optimization of spatial graphs. Then, the Temporal Extension Module [17] is applied for extraction of features from the temporal dimension, which not only can process the same joints between frames but also can pay attention to multiple adjacent joints between frames. This is helpful to ensure extraction of more abundant temporal features. Furthermore, the attention mechanism [18] is introduced to obtain more important information of joints, frames and channels and remove redundant feature information, thus further improving the performance of our method.

The contributions of this paper are summarized as follows:

i. A video behavior recognition method is proposed based on Actional-Structural Graph Convolution and a Temporal Extension Module, which simultaneously optimizes the spatial and temporal features.

ii. In the spatial dimension, Actional-Structural Graph Convolution is composed of two networks: action graph convolution and structural graph convolution. Among them, the action graph convolution extracts rich spatial features by capturing the correlations between distant joint features, whereas the structural graph convolution extends the existing skeleton graphs to obtain the spatial features of multiple adjacent joints.

iii. In the time dimension, the Temporal Extension Module is introduced. The conventional methods only gain the same joint features of adjacent frames. Nevertheless, our method acquires the joint features of the same position and adjacent positions in adjacent frames, which expands on the temporal graphs to extract more abundant temporal features.

iv. A large number of experiments are carried out on two standard behavior recognition datasets to evaluate the effectiveness and feasibility of our proposed method. Compared with some existing behavior recognition methods, the experimental results show that our method can achieve better results.

2. Related work

The complexity of human posture, occlusion and illumination affects behavior recognition results. With the application of depth cameras and the progress of posture estimation algorithms, skeletal data has been proved to be an effective source of behavior information. Moreover, skeletal data minimizes the effects of irrelevant factors such as occlusion, illumination and human clothing on behavior recognition in RGB images. Therefore, some scholars are committed to combining deep learning methods with skeletal data to improve behavior recognition. Generally, there are three kinds of neural networks that are used for skeleton-based behavior recognition: Convolution Neural Networks (CNNs), Recurrent Neural Networks (RNNs) and Graph Convolution Neural Networks (GCNs).

2.1. Skeleton-based behavior recognition methods using CNNs

Skeleton-based behavior recognition methods using CNNs mainly convert three-dimensional skeletal data into pseudo images for processing. Kim and Reiter [19] used a class of models known as Temporal Convolutional Neural Networks (TCNs) to explicitly learn readily interpretable spatio-temporal representations for 3D human action recognition. Ke et al. [20] introduced a method for 3D action recognition with skeleton sequences, which transformed each skeleton sequence into three clips consisting of several frames for spatial temporal feature learning using deep neural networks. Liu et al. [21] presented an enhanced skeleton visualization method for view-invariant human action recognition based on CNNs. Li et al. [22] designed a multi-scale dilated convolutional neural network for the classification of skeleton images. Hu et al. [23] systemized behavior recognition algorithms based on deep learning in recent years. These methods based on CNNs are relatively simple, and some light CNN models are fast and less computationally intensive. However, the models have low accuracy, and they cannot effectively filter the background noise in the data.

2.2. Skeleton-based behavior recognition methods using RNNs

Skeleton-based behavior recognition methods using RNNs extract sequential information to solve sequence problems. A RNN has a memory module that can utilize the temporal information, and thus the vertices of the network can be effectively connected. Liu et al. [24] extended RNNs in the spatial domain and the temporal domain to better analyze the hidden sources of action-related information within the human skeleton sequences in both of the two domains simultaneously. Liu et al. [25] proposed the Global Context-Aware Attention LSTM (GCA-LSTM) for 3D action recognition, which focused on the informative joints in the action sequence with the assistance of global contextual information. Wang et al. [26] presented a video architecture, termed as Temporal Difference Network (TDN), which mainly captured multi-scale temporal information for effective action recognition. Liu et al. [27] merged the results of the motion history images input into VGG-16 and the RGB image input into the Faster R-CNN algorithm for human abnormal behavior recognition. Si et al. [28] introduced a model with spatial reasoning and temporal stack learning (SR-TSL) for skeleton-based action recognition, which consisted of a spatial reasoning network (SRN) and a temporal stack learning network (TSLN). The methods based on RNNs can effectively maintain the context information and have high recognition accuracy. However, they have limitations such as gradient explosion and gradient disappearance, and they cannot eliminate the redundant information of joint data.

2.3. Skeleton-based behavior recognition methods using GCNs

A graph is composed of vertices and edges. The feature information of each vertex in the graph is independent, and any two vertices in the graph may have a relationship with irregular data structures. Because graph data does not have translation invariance, it is challenging to apply GCNs to skeletal data. Yang et al. [29] proposed a channel adaptive merging module specific for the human skeleton graph, which can adaptively and efficiently merge the vertices from the same part of the skeleton graph. Chen et al. [30] presented a multi-scale spatial graph convolution (MS-GC) module and a multi-scale temporal graph convolution (MT-GC) module to enhance the receptive field of the model in spatial and temporal dimensions. Ding et al. [31] put forward a temporal segment graph convolutional network (TS-GCN) for skeleton-based action recognition. Shi et al. [32] raised a two-stream adaptive graph convolutional network (2s-AGCN) for skeleton-based action recognition. Zhang et al. [33] designed a simple but effective semantics-guided neural network (SGN) for skeleton-based action recognition, which explicitly introduced the high-level semantics of joints into the network to enhance the feature representation capability. Si et al. [34] gave an Attention Enhanced Graph Convolutional LSTM Network (AGC-LSTM) for human action recognition based on skeleton data, which can capture discriminative features in spatial configuration and temporal dynamics. Miao et al. [35] proposed a graph convolutional operator referred to as a central difference graph convolution (CDGC) for skeleton-based action recognition, which aggregated node information and gradient information similar to a vanilla graph convolutional operation. Chen et al. [36] presented a Channel-wise Topology Refinement Graph Convolution (CTR-GC) to dynamically learn different topologies and effectively aggregate joint features in different channels for skeleton-based action recognition.

GCNs have been widely used for skeleton-based behavior recognition due to their excellent modeling ability for non-Euclidean data. Because skeletal data is embedded in the form of graphs, rather than vectors or images, RNNs and CNNs have poor representation of skeletal data compared with GCNs. As the graph convolution is a local operation, it can only utilize the short-range joint dependencies and short-term trajectory, and it can fail to directly model the distant joints' relations and long-range temporal information that are essential in distinguishing various behaviors. Furthermore, the fixed network would cause many redundancies in the representation of behaviors and deteriorate the performance. Moreover, the redundant features may hinder the model from focusing on significant features. To mitigate the above issues, this paper puts forward a novel method to extract the behavior features from different joints, frames and channels, which is essential for skeleton-based behavior recognition in videos.

3. The proposed method

Recently, skeleton-based behavior recognition has made great progress, but many problems still remain unsolved. For example, the representations of skeleton sequences captured by most of the previous methods lack spatial structure information and detailed temporal dynamics features. Consequently, we utilize Actional-Structural Graph Convolution to get rich spatial features by capturing the correlations of distant joint features and extend the existing skeleton graph to obtain the spatial features of multiple adjacent joints. Then, the Temporal Extension Module (TEM) is used for gaining the joint features of the same position and adjacent positions in adjacent frames. Meanwhile, we apply three attention mechanisms to improve the performance of video behavior recognition.

The architecture of our method is shown in Figure 1. The proposed method takes the skeletal data as the input of the whole network. Our basic network structure is composed of nine basic units, each of which is successively composed of a spatial graph convolution module, an attention mechanism module and a temporal extension module. In order to learn spatial and temporal features more effectively from important joints, frames and channels, we integrate the above three modules into a network structure.

Specifically, we apply nine layers of integrated modules in total. Taking one layer network for an example, the skeletal data is first input into the spatial graph convolution, consisting of action graph convolution and structural graph convolution. After that, the features are processed through a batch normalization (BN) layer and ReLU activation function. Then, the attention mechanism is adopted to focus on important frames, joints and channels for realizing the optimization of features, which includes three parts: spatial attention, temporal attention and channel attention. Finally, the TEM can further extract the temporal features. Like the spatial graph convolution, this module also processes the features through the BN and ReLU activation function. Behind the whole backbone network, the features go through the Global Ave-Pooling layer, and the behavior recognition results are obtained through the Softmax layer.

3.1. Spatial graph convolution

The spatial graph convolution used in this paper is improved on the basis of spatio-temporal graph convolution. A graph $G = (V, E)$ is constructed based on a skeletal sequence, where V represents the set of all vertices, and E includes two parts: action graph convolution and structural graph convolution. In this way, the features of distant joints and adjacent joints can be extracted to enrich the features of the spatial dimensions.

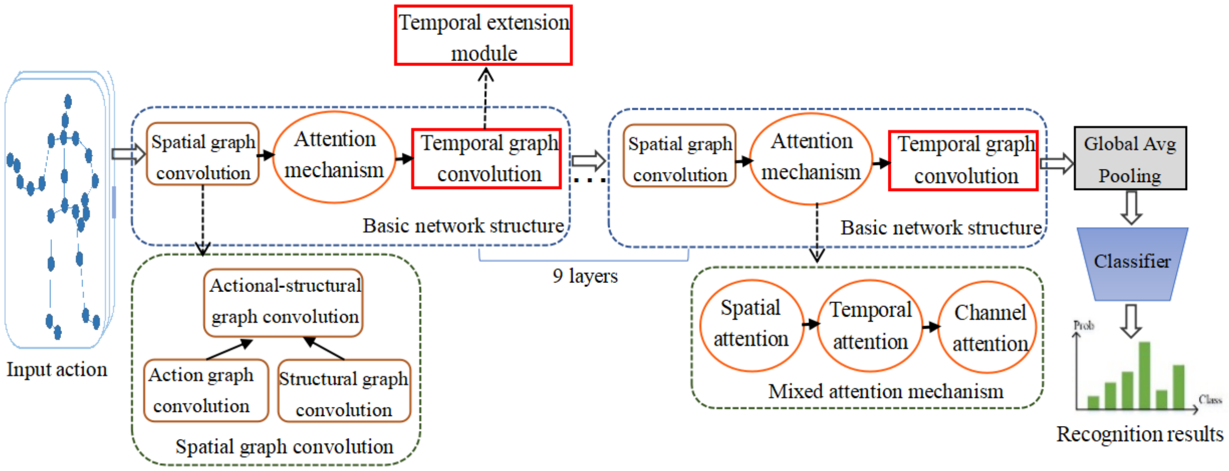

Temporal extension module
Input action
Spatial graph convolution
Attention mechanism
Temporal graph convolution
Basic network structure
9 layers
Spatial graph convolution
Attention mechanism
Temporal graph convolution
Basic network structure
Global Avg Pooling
Classifier
Actional-structural graph convolution
Action graph convolution
Structural graph convolution
Spatial graph convolution
Spatial attention
Temporal attention
Channel attention
Mixed attention mechanism
Prob
Class
Recognition results

Figure 1. The architecture of our proposed method.

3.1.1. Action graph convolution

The encoder-decoder structure [37] of the Neural Relational Inference (NRI) model is used to realize the message passing between distant joints, as shown in Figure 2. The encoder takes the form of a GNN with multiple rounds of node-to-edge ($v \rightarrow e$) and edge-to-node ($e \rightarrow v$) message passing to get the relevance between distant joints. According to the connections of distant joints obtained by the encoder, the decoder runs multiple GNNs in parallel for extracting the features of joints. The features obtained by action connection are convoluted with the convolution kernel, which is called action graph convolution. Action graph convolution realizes message passing between distant joints.

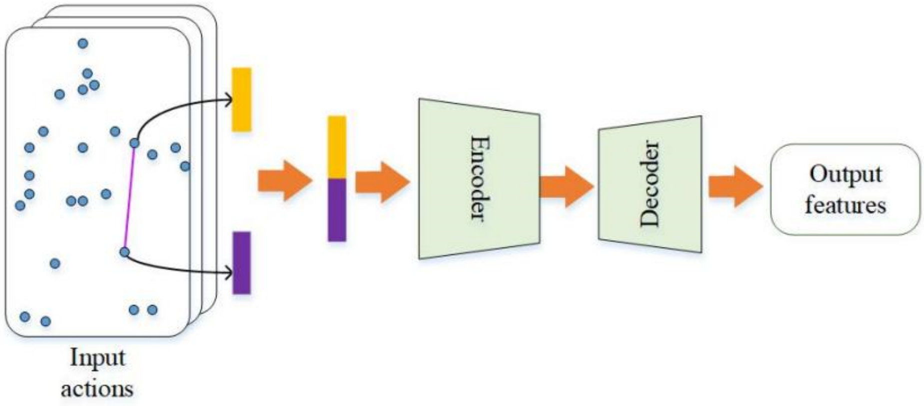


Figure 2. The structure of encoder-decoder. The yellow rectangle represents the aggregation of joint features, and the purple rectangle represents the connection features between joints. The two are connected in series to obtain the joint feature vector and input it to the encoder.

The function of the encoder is to acquire the message passing between joints according to the skeleton graph, which is calculated by Eqs (1) to (3). Eq (1) represents the connection features between joints, Eq (2) represents the aggregation of joint features, and Eq (3) represents the probability of joint connection.

$$Q_{\mathrm{i},j}^{(k+1)} = f_e^{(k)}(f_v^{(k)}(p_i^{(k)}) \oplus (f_v^{(k)}(p_j^{(k)}))) \tag{1}$$

$$\mathrm{p}_{\mathrm{i}}^{(\mathrm{k}+1)} = F(Q_{i,:}^{(k+1)}) \oplus p_i^{(k)} \tag{2}$$

$$A_{i,j:} = soft\max(\frac{Q_{i,j}^{(K)} + r}{\tau}) \in R^C \tag{3}$$

where $f_e(\cdot)$ denotes the Multi Layer Perceptron acting on edges, $f_v(\cdot)$ denotes the Multi Layer Perceptron acting on joints, P_i denotes the *i*-th feature of the joint, $\oplus$ is vector splicing, *k* represents the number of iterations, and $Q_{i,j}$ indicates the connection feature of joint *i* and joint *j*. $F(\cdot)$ denotes the aggregation operation. *r* stands for a random vector, and τ is used to control the discretization of probability.

The connection features between joints are aggregated and then spliced with the original features to gain the features after message passing. The function of the decoder is to extract joint features according to the connection probability between joints obtained by the encoder. The decoder is calculated by Eqs (4) and (5):

$$Q_{i,j}^t = \sum\nolimits_{c=1}^{C} A_{i,j,c} f_e^{(c)}(f_v^{(c)}(x_i^t) \oplus f_v^{(c)}(x_j^t)) \tag{4}$$

$$p_i^t = F(Q_{i,:}^t) \oplus p_i^t \tag{5}$$

where *C* represents the number of connections between joints, $A_{i,j,c}$ denotes the connection probability between joints obtained by the encoder, and *t* represents *t*-th frame. The encoder-decoder operation can capture the dependence between distant joints.

3.1.2. Structural graph convolution

The natural connection matrix of the existing methods is extended by high-order form. Existing methods only focus on one adjacent joint and ignore the feature relationship with other adjacent joints. Fortunately, the information of multiple adjacent joints can be obtained through structural graph convolution. Taking the joint features obtained by structural connection as the convolution kernel for the convolution operation is structural graph convolution. The calculation of structural graph convolution is shown in Eq (6):

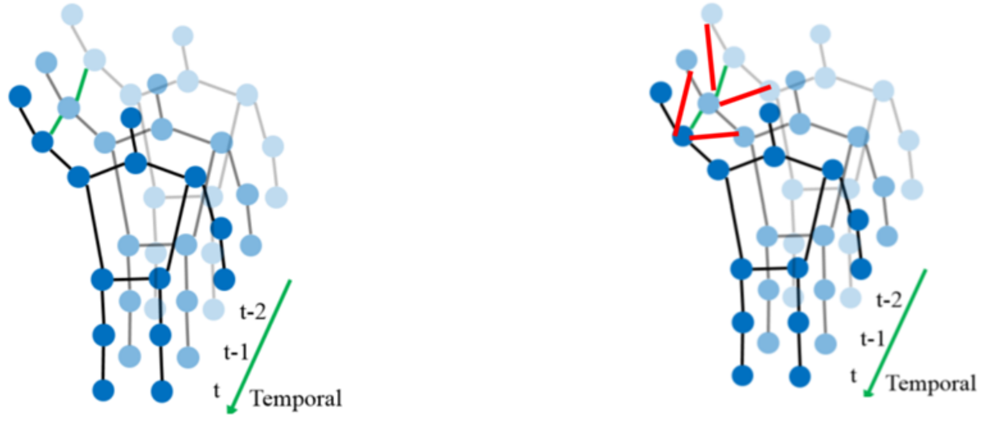
$$X_{out} = \sum_{l=1}^{L} \sum_{p \in P} M^{(l)} \circ A_1^{(p,l)} X_{in} W^{(p,l)}, \tag{6}$$

where $A_1 = D^{-1}A$, A is the adjacency matrix, D is the stiffness matrix of A for normalization, X_{in} represents the input features of the joint, l represents the numbers of adjacent joints, $M^{(l)}$ denotes trainable weights of edges, $W^{(p,l)}$ denotes trainable weights to obtain the importance of features, and $\circ$ is the Hadamard product to realize the multiplication of corresponding matrix elements.

Furthermore, action graphs and structural graphs are linked to realize the extraction of spatial features. Because action connection only considers the coordinates of joints, while structural connection only considers the connection of joints in physical structure, their contributions do not affect each other and meet the linear relationship. Through the message passing of distant joints and multiple adjacent joints, Actional-Structural Graph Convolution enriches the spatial features and improves the performance of the model.

3.2. Temporal graph convolution

Most of the existing methods concentrate upon optimizing the spatial graph but ignore the optimization of the temporal graph. Temporal modeling is still a challenge for behavior recognition in videos. The traditional temporal graph is shown in Figure 3(a). The blue spots in the figure indicate the joints, the black lines represent the connections of joints in the spatial dimension, and the green lines show the connection in the time dimension. It can be seen that the traditional temporal graphs only concatenate the same joints of adjacent frames in the time dimension, without the information of adjacent joints in adjacent frames. They generally use the simple convolution network, which cannot provide abundant temporal features.



(a) The traditional temporal graph.

(b) The improved temporal graph.

Figure 3. The differences of traditional temporal graph and improved temporal graph.

Hence, we introduce the Temporal Extension Module, as shown in Figure 3(b). The green lines in the figure represent the connections of the same joints in adjacent frames, and the red lines represent the connections of adjacent joints in adjacent frames. In the time dimension, we expand the sampling range to achieve message passing of adjacent joints in adjacent frames. The calculation of TEM is shown in Eqs (7) and (8):

$$f_{out}(v_{ti}) = \sum_{v(t-1)j \in B^T(v_{ti})} \frac{1}{Z_{ti}(v_{(t-1)j})} f_{in}(v_{(t-1)j}) \cdot w(l_{(t-1)i}(v_{(t-1)j})) \tag{7}$$

$$B^T(v_{ti}) = \{v_{(t-1)j} \middle| d(v_{(t-1)j}, v_{(t-1)i}) \le D^T\} \tag{8}$$

where $l_{(t-1)i}(v_{(t-1)j})$ denotes the label mapping of node *i* relative to node *j* in the *t*-1 frame, $f_{in}(v_{(t-1)j})$ represents the input feature of node *j* in the *t*-1 frame, and $w(\cdot)$ indicates the weight vector. $B^T(v_{ti})$ represents sampling range, $d(v_{(t-1)j}, v_{(t-1)i})$ is the minimum length of the path from the node v_i to v_j, and D^T denotes the maximum length of inter-frame sampling. When D^T =1, it means only one adjacent node.

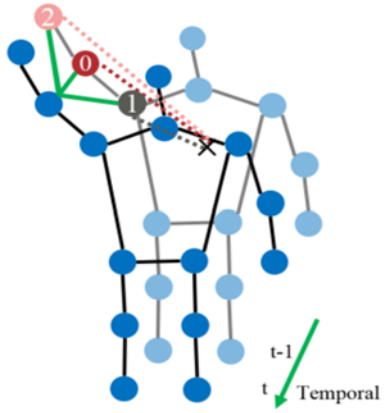


Figure 4. Extended sampling range.

The sample set is divided into three subsets, as shown in Figure 4. The black cross in the figure shows the center of gravity of the whole skeleton. The same joint as the previous frame is the root node, which is marked as 0. The adjacency node whose distance to the center of gravity is closer than the root node represents the centripetal motion feature, which is marked as 1. The adjacency node whose distance to the center of gravity is farther than the root node represents the centrifugal motion feature, which is marked as 2. Extended sampling range is conducive to extracting discriminative temporal features for improving the accuracy of behavior recognition.

3.3. Attention mechanism

To explore the internal relationships of the data and highlight the important features, three types of attention mechanisms are applied as attention mechanism modules for skeleton-based behavior recognition, as shown in Figure 5. This module includes three parts: spatial attention, temporal attention and channel attention.

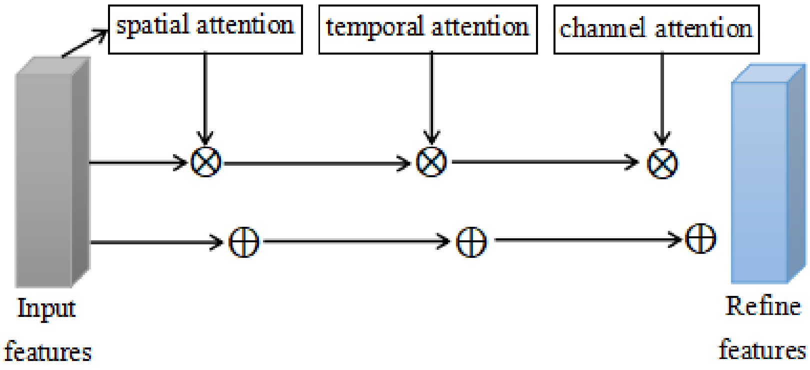


Figure 5. The network structure of attention mechanism.

The spatial attention module assigns different weights according to the participation of each joint and extracts more important joint information. The calculation is shown in Eq (9):

$$M_s = \sigma(g_s(AvgPool(f_{in}))) \tag{9}$$

where f_{in} represents input features, $AvgPool(\cdot)$ indicates the average pooling of features, g_s is a one-dimensional convolution operation, σ denotes the Sigmoid activation function, and M_s represents attention mapping. The input features are multiplied by M_s in the form of residuals to refine the joint features, and the weights of different joints are adaptively adjusted by training the network. Similar to the spatial attention, the calculations of temporal attention and channel attention are shown in Eqs (10) and (11):

$$M_t = \sigma(g_t(AvgPool(f_{in}))) \tag{10}$$

$$M_{\mathrm{c}} = \sigma(W_2(\delta(W_1(AvgPool(f_{in}))))) \tag{11}$$

Our attention mechanism is a typically mixed attention mechanism that can first infer the attention map along the three dimensions of spatial, temporal and channel and then weight the input feature map by the attention map to complete the adaptive feature optimization.

3.4. Implementation details

We implement the experiment by the PyTorch library. Our basic network structure is composed of nine basic units, each of which is successively composed of a spatial graph convolution module, an attention mechanism module and a temporal extension module. Among the nine units, the feature dimensions are 64, 64, 64, 128, 128, 128, 256, 256, 256, sequentially. The strides of convolution are 1, 1, 1, 2, 1, 1, 2, 1, 1, respectively. In order to optimize our network, the stochastic gradient descent algorithm (SGD) is selected as the optimization function, and the initial momentum is set to 0.9.

In the process of training, we contrast the output of the Softmax classifier to the original label and update the parameters by error back propagation. The probability of back propagation is set to 0.5. Cross entropy loss is used as the loss function. The configuration of the computer is as follows: The

CPU is an Intel Core i7-9700K, and the VGA is an NVIDIA GeForce GTX1080Ti. This shows that our method can complete the experiment without GPU, which reduces computing costs and requirements.

4. Experiments and result analysis

In this section, we execute experiments to demonstrate the performances of the Actional-Structural Graph Convolution, Attention Mechanisms and Temporal Extension Module, respectively. Furthermore, the proposed method is compared with superior behavior recognition methods on two standard datasets: NTU-RGB+D and Kinetics.

4.1. Experimental dataset

The behavior recognition datasets NTU-RGB+D [38] and Kinetics [39] were used to train and test our method. NTU-RGB+D contains 60 different action classes, including daily, mutual and health-related actions. The dataset consists of 56,880 RGB+D video samples, captured from 40 different human subjects, using a Microsoft Kinect v2 sensor. Four major data modalities are provided by this sensor: depth maps, 3D joint information, RGB frames and IR sequences. In this paper, we only use the 3D joint information. Joint information consists of 3-dimensional locations of 25 major body joints. The configuration of body joints is illustrated in Figure 6.

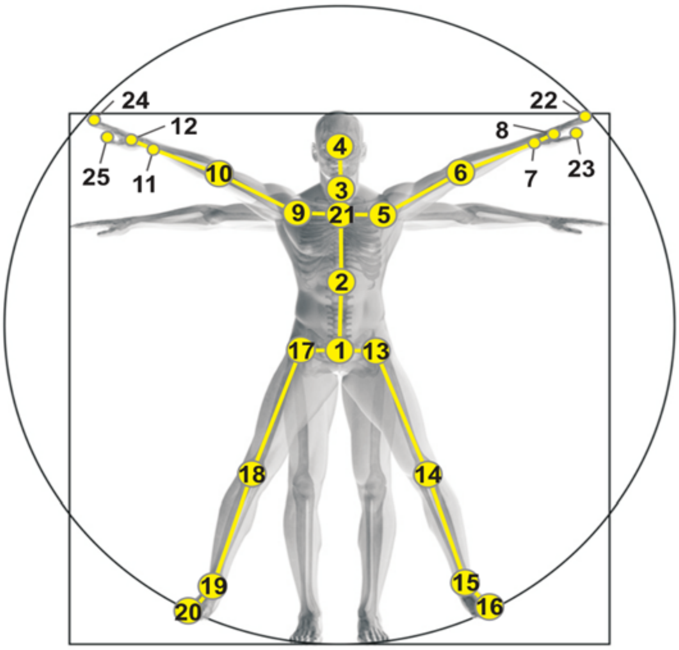


Figure 6. Configuration of 25 body joints in NTU-RGB+D.

Two standards are served as performance evaluation in NTU-RGB+D: cross-subject (CS) and cross-view (CV). In cross-subject evaluation, they split the 40 subjects into training and testing groups. Each group consists of 20 subjects, and the training and testing sets incorporate 40, 320 and 16,560 samples, respectively. For cross-view evaluation, all the samples of camera 1 are for testing, and samples of cameras 2 and 3 are for training. The training and testing sets contain 37,920 and 18,960 samples, respectively. Sample frames of the NTU-RGB+D dataset are illustrated in Figure 7.

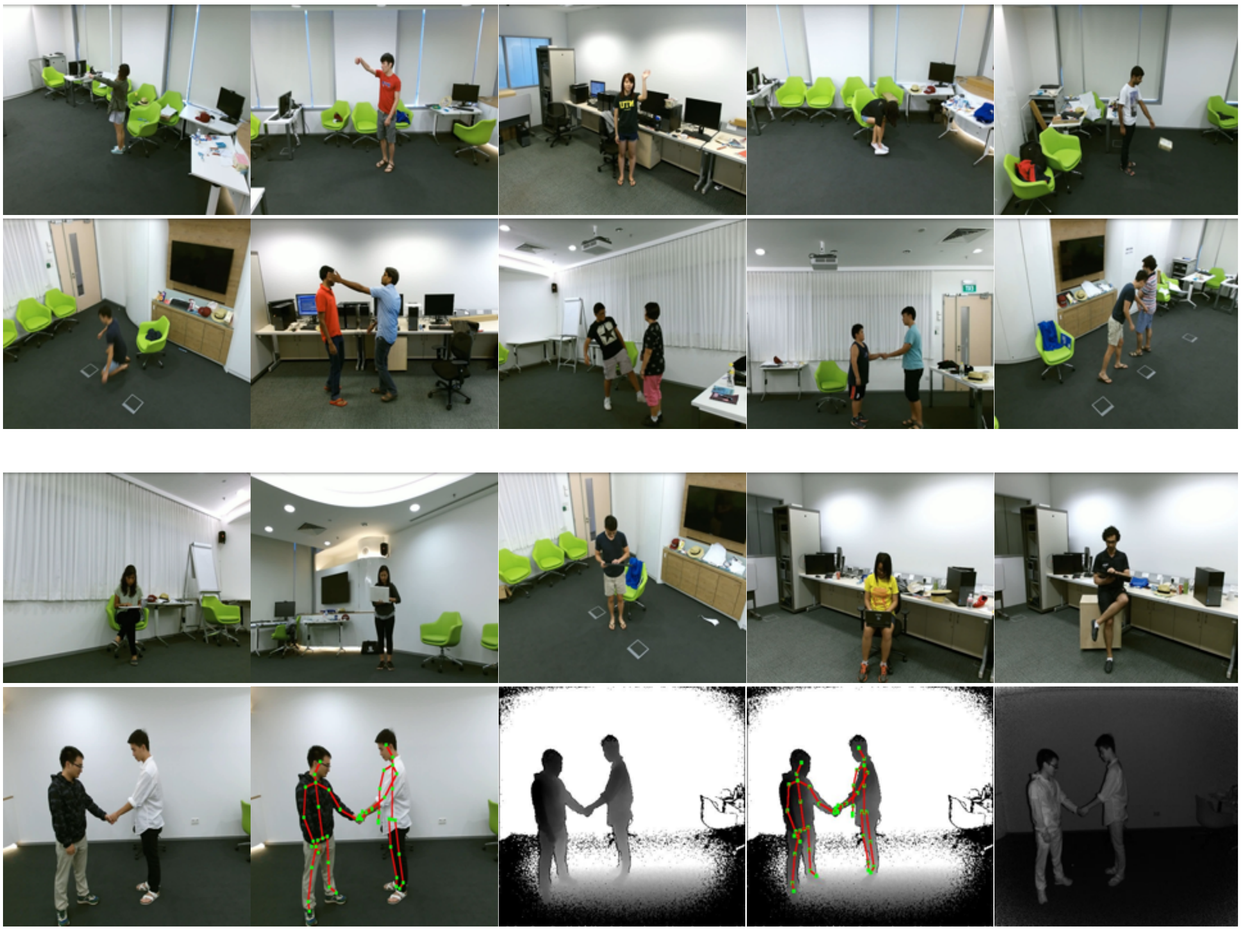


Figure 7. Sample frames of the NTU-RGB+D dataset. The second and third rows show the varieties in human subjects and camera views, respectively. The last row illustrates RGB, RGB + joints, depth, depth + joints and IR modalities of a sample frame.

Kinetics provides a large scale high quality dataset. The dataset has 400 human action classes, with 400–1150 clips for each action, each from a unique video. Each clip lasts around 10 s. The current version includes 306,245 videos. The actions are human focused and cover a broad range of classes, including human-object interactions as well as human-human interactions. This dataset supplies the original video clips. In our experiment, the public OpenPose toolbox is used to extract the positions of 18 joints in each frame, and the joint labels are from 0 to 17, as shown in Figure 8.

The training sets and testing sets of the Kinetics dataset are 240,000 video clips and 200,000 video clips, respectively. In the testing sets, top-1 and top-5 are used to evaluate the performance of the method. Top-1 takes the corresponding action with the highest probability of classification as the prediction result, and top-5 shows the top five actions with the highest probability of classification. When the ground truth matches one of the top five, the prediction is correct. Example classes from the Kinetics dataset are shown in Figure 9.

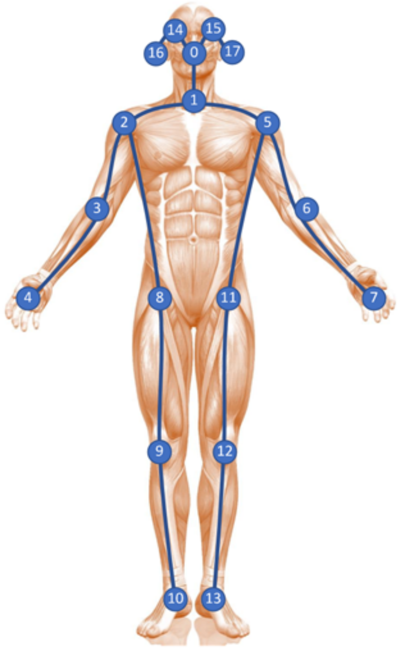
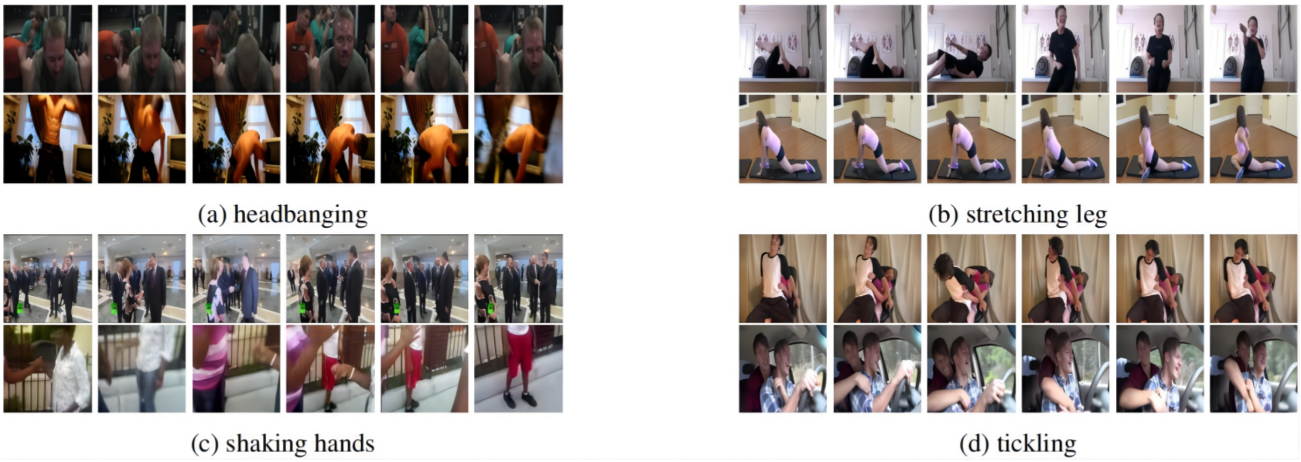


Figure 8. Configuration of 18 joints extracted in Kinetics.


(a) headbanging

(b) stretching leg

(c) shaking hands

(d) tickling

Figure 9. Example classes from the Kinetics dataset.

4.2. Experiments and result analysis

Compared with the latest behavior recognition method on the CS standard and CV standard of the NTU-RGB+D dataset, the results are shown in Table 1. [8] extracted features manually, [36,40] adopted the methods based on RNNs, [41–44] employed the methods based on CNNs, and [16,28,45–47] used the methods based on GCNs. It can be seen from Table 1 that our method has improved the accuracy of behavior recognition, whose performance is better than the existing methods. This is because our method can not only capture the message passing of distant joints in the same frame and adjacent joints in different frames but also extract more important joint, frame and channel information, which is helpful to improve the performance of behavior recognition.

Table 1. Comparison results on NTU-RGB+D dataset.

Method	CS	CV
Lie Group [8]	50.1%	52.8%
H-RNN [40]	59.1%	64.0%
Deep LSTM [38]	60.7%	67.3%
PA-LSTM [38]	62.9%	70.3%
Two-Stream 3DCNN [41]	66.8%	72.6%
Visualize CNN [42]	76.0%	82.6%
CNN + Motion + Trans [44]	83.2%	89.3%
3scale ResNet152 [44]	85.0%	92.3%
STGCN [45]	81.5%	88.3%
SDGCN [46]	84.0%	91.4%
SR-TSL [28]	84.8%	92.4%
AS-GCN [16]	86.8%	94.2%
RA-GCNv2 [47]	87.3%	93.6%
SAR-NAS [48]	86.4%	94.3%
TS-SAN [49]	87.2%	92.7%
MANs [50]	82.7%	93.2%
Our method	87.5%	94.7%

In addition, the comparative experiments on the Kinetics dataset are shown in Table 2. [51] is based on handmade feature extraction, [16,45] used the methods based on GCNs, and [19] utilized the methods based on CNNs. From Table 2, we can conclude that the accuracy of our method on top-1 and top-5 is higher than the existing methods, which verifies the effectiveness of the behavior recognition method based on Actional-Structural Graph convolution and TEM proposed in this paper.

Table 2. Comparison results on Kinetics dataset.

Method	top-1	top-5
Feature Enc [51]	14.9%	25.8%
Deep LSTM [38]	16.4%	35.3%
TCN [19]	20.3%	40.0%
STGCN [45]	30.7%	52.8%
AS-GCN [16]	34.8%	56.5%
Our method	35.7%	57.8%

4.3. Ablation experiment

Ablation experiments were carried out on the NTU-RGB+D dataset. In this experiment, the performances of spatial graph convolution, attention mechanism and temporal extension module in our network architecture were proved, respectively.

4.3.1. Spatial graph convolution

The availability of action graph convolution and structural graph convolution are validated. First, we contrast the method with action graph convolution to the behavior recognition method based on STGCN. The comparison results of action graph convolution are shown in Table 3. The experiment shows that the accuracy of the method with action graph convolution is higher, which means that capturing the dependence between distant joints makes for behavior recognition.

Table 3. The results of adding action graph convolution.

Method	CS	CV
STGCN	81.5%	88.3%
STGCN + Action	83.2%	90.3%

Second, we verify the effectiveness of structural graph convolution, as shown in Table 4. The structural graph convolution introduced in this paper takes the STGCN as the baseline framework. In this experiment, the number of capturing adjacent joints is set to 1–5 on the spatial graph. When it is 1, the corresponding structural graph convolution is the same as the STGCN. From Table 4, we can find that as the order increases, the accuracy of recognition will improve, which shows that capturing the message passing of multiple adjacent joints is conducive to behavior recognition. However, when the order is 5, the accuracy begins to decrease, because with the increase of the order, the method cannot obtain local features well.

Table 4. The results of adding structural graph convolution.

Method	CS	CV
STGCN	81.5%	88.3%
STGCN + 1-order Structure	81.5%	88.3%
STGCN + 2-order Structure	82.2%	89.1%
STGCN + 3-order Structure	83.4%	89.6%
STGCN + 4-order Structure	84.2%	90.2%
STGCN + 5-order Structure	83.5%	90.1%

Finally, we test the effect of combining action graph convolution and structural graph convolution, as shown in Table 5. It can be seen from Table 5 that when the order of structural connection is 4, the performance of the model is best. Therefore, the spatial graph convolution in this paper uses action graph convolution + 4-order structural graph convolution, named STGCN-AS in the following.

Table 5. The results of spatial graph convolution.

Method	CS	CV
STGCN	81.5%	88.3%
STGCN + 1-order Structure + Action	83.2%	90.3%
STGCN + 2-order Structure + Action	83.7%	91.2%

Continued on next page

Method	CS	CV
STGCN + 3-order Structure + Action	84.4%	92.3%
STGCN + 4-order Structure + Action (STGCN-AS)	86.1%	93.2%
STGCN + 5-order Structure + Action	84.2%	92.0%

4.3.2. Attention mechanism

The significances of three attention mechanisms are confirmed in the framework of STGCN-AS, respectively, as shown in Table 6. The experimental results show that the three attention mechanisms are all useful for improving the performance of the method, which achieves the best results when the mixed attention mechanism is added. That is, collecting the features from important joints, frames and channels is beneficial for behavior recognition.

Table 6. The results of adding attention mechanisms.

Method	CS	CV
STGCN-AS	86.1%	93.2%
STGCN-AS + Spatial Attention	86.4%	93.3%
STGCN-AS + Temporal Attention	86.2%	93.2%
STGCN-AS + Channel Attention	86.3%	93.4%
STGCN-AS + Mixed Attention	86.5%	93.5%

4.3.3. Temporal extension module

The temporal extension module based on STGCN-AS was verified, as shown in Table 7. The results show that adding the temporal extension module can enhance the performance of the method, because this module can expand the range of feature extraction in the time dimension and enrich the temporal features.

Table 7. The results of adding temporal extension module.

Method	CS	CV
STGCN-AS	86.1%	93.2%
STGCN-AS + TEM	86.7%	93.8%

4.3.4. Overall model

The above experiments have proved the applicability of the three modules proposed in this paper. The overall model of this paper is arranged according to the order of spatial graph convolution, attention mechanism and temporal extension module. The experimental results are shown in Table 8. It can be seen from Table 8 that optimizing the spatio-temporal map and paying attention to important joints, frames and channels are beneficial for improving the performance of behavior recognition. When adding one of three modules, the accuracy is improved slightly, but the accuracy of the overall model reaches an improvement of 1.4–1.5% in terms of both the CS and CV standards compared with the baseline method.

Table 8. The results of the overall model.

Method	CS	CV
STGCN-AS	86.1%	93.2%
STGCN-AS + Attention	86.5%	93.5%
STGCN-AS + TEM	86.7%	93.8%
STGCN-AS + Attention + TEM	87.5%	94.7%

In summary, skeleton-based action recognition is an important task that requires the adequate understanding of movement characteristics of a human action from the given skeleton sequence. Hence, we propose a novel method based on STGCN-AS and a temporal extension module to extract the abundant features of different joints, frames and channels, which is significant for video behavior recognition. The effectiveness of the proposed method is verified by a large number of comparative experiments, and the performances of the three modules are demonstrated by a series of ablation experiments, respectively.

5. Conclusions and future work

The GCN has shown promising performance for behavior recognition due to its strengths in modeling the dependencies and dynamics in sequential data. Therefore, we propose a novel method based on STGCN-AS and a temporal extension module to extract the abundant features of different joints, frames and channels. STGCN-AS is utilized to get rich spatial features by capturing the correlations of distant joint features, and the existing skeleton graphs are extended to obtain the spatial features of multiple adjacent joints. Then, the TEM is used for gaining the joint features of the same position and adjacent positions in adjacent frames. Meanwhile, three attention mechanisms are applied to improve the performance of video behavior recognition. Although our method has achieved good recognition accuracy on the NTU-RGB+D and Kinetics datasets, there are still deficiencies. Facial expressions, gestures and other features are not involved in our method. In the future, we will combine gesture recognition with behavior recognition to further improve the performance of our method and research the features of multi-person interaction.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61907007), Fund of Jilin Provincial Science and Technology Department (20210201077GX).

Conflict of interest

The authors declare no conflict of interest.

References

1. J. K. Aggarwal, M. S. Ryoo, Human activity analysis: A review, *ACM Comput. Surv.*, **43** (2011), 1–43. https://doi.org/10.1145/1922649.1922653

2. H. Wang, C. Schmid, Action recognition with improved trajectories action recognition with improved trajectories, in *2013 IEEE International Conference on Computer Vision*, IEEE, Sydney, NSW, Australia, (2013), 3551–3558. https://doi.org/10.1109/ICCV.2013.441
3. Y. H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, G. Toderici, Beyond short snippets: Deep networks for video classification, in *2015 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, (2015), 4694–4702. https://doi.org/10.1109/CVPR.2015.7299101
4. Z. Qin, Y. Liu, M. Perera, S. Anwar, T. Gedeon, P. Ji, et al., ANUBIS: Review and benchmark skeleton-based action recognition methods with a new dataset, preprint, arXiv:2205.02071.
5. Z. Zhang, Y. Hu, S. Chan, L. T. Chia, Motion context: A new representation for human action recognition, in *European Conference on Computer Vision*, Academic press, (2008), 817–829. https://doi.org/10.1007/978-3-540-88693-8_60
6. J. C. Niebles, H. Wang, F. F. Li, Unsupervised learning of human action categories using spatial-temporal words, *Int. J. Comput. Vision*, **79** (2008), 299–318. https://doi.org/10.1007/s11263-007-0122-4
7. H. Wang, A. Klser, C. Schmid, C. L. Liu, Dense trajectories and motion boundary descriptors for action recognition, *Int. J. Comput. Vision*, **103** (2013), 60–79. https://doi.org/10.1007/s11263-012-0594-8
8. R. Vemulapalli, F. Arrate, R. Chellappa, Human action recognition by representing 3D skeletons as points in a lie group, in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Columbus, USA, (2014), 588–595. https://doi.org/10.1109/CVPR.2014.82
9. M. E. Hussein, M. Torki, M. A. Gowayyed, M. A. El-Saban, Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations, in *Twenty-third International Joint Conference on Artificial Intelligence*, AAAI, Beijing, China, (2013), 2466–2472.
10. F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, R. Bajcsy, Sequence of the most informative joints (smij): A new representation for human skeletal action recognition, *J. Visual Commun. Image Represent.*, **25** (2014), 24–38. https://doi.org/10.1016/j.jvcir.2013.04.007
11. L. Xia, C. C. Chen, J. K. Aggarwal, View invariant human action recognition using histograms of 3D joints, in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, IEEE, Providence, USA, (2012), 20–27. https://doi.org/10.1109/CVPRW.2012.6239233
12. C. Li, Q. Zhong, D. Xie, S. Pu, Skeleton-based action recognition with convolutional neural networks, in *2017 IEEE International Conference on Multimedia & Expo Workshops*, IEEE, Hong Kong, (2017), 597–600. https://doi.org/10.1109/ICMEW.2017.8026285
13. C. Li, Q. Zhong, D. Xie, 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*, AAAI, Stockholm, Sweden, (2018), 786–792. https://doi.org/10.24963/ijcai.2018/109
14. C. Caetano, J. Sena, F. Bremond, J. A. Dos Santos, W. R. Schwartz, Skelemotion: A new representation of skeleton joint sequences based on motion information for 3D action recognition, in *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance,* IEEE, Taipei, Taiwan, (2019), 1–8. https://doi.org/10.1109/AVSS.2019.8909840

15. Y. Li, R. Xia, X. Liu, Q. Huang, Learning shape-motion representations from geometric algebra spatio-temporal model for skeleton-based action recognition, in *2019 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, Shanghai, China, (2019), 1066–1071. https://doi.org/10.1109/ICME.2019.00187
16. M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, Q. Tian, Actional-structural graph convolutional networks for skeleton-based action recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, (2019), 3595–3603. https://doi.org/10.1109/CVPR.2019.00371
17. S. Song, C. Lan, J. Xing, W. Zeng, J. Liu, An end-to-end spatio-temporal attention model for human action recognition from skeleton data, in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI, San Francisco, USA, (2016), 4263–4270. https://doi.org/10.1609/aaai.v31i1.11212
18. L. Shi, Y. Zhang, J. Cheng, H. Lu, Skeleton-based action recognition with multi-stream adaptive graph convolutional networks, in *IEEE Transactions on Image Processing*, IEEE, (2020), 9532–9545. https://doi.org/10.1109/TIP.2020.3028207
19. T. S. Kim, A. Reiter, Interpretable 3D human action analysis with temporal convolutional networks, in *2017 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Honolulu, USA, (2017), 1623–1631. https://doi.org/10.1109/CVPRW.2017.207
20. Q. Ke, M. Bennamoun, S. An, F. Sohel, F. Boussaid, A new representation of skeleton sequences for 3D action recognition, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Honolulu, USA, (2017), 4570–4579. https://doi.org/10.1109/CVPR.2017.486
21. M. Liu, L. Hong, C. Chen, Enhanced skeleton visualization for view invariant human action recognition, *Pattern Recognit.*, **68** (2017), 346–362. https://doi.org/10.1016/j.patcog.2017.02.030
22. B. Li, M. He, Y. Dai, X. Cheng, Y. Chen, 3D skeleton based action recognition by video-domain translation-scale invariant mapping and multi-scale dilated CNN, *Multimed. Tools Appl.*, **77** (2018), 22901–22921. https://doi.org/10.1007/s11042-018-5642-0
23. K. Hu, J. Jin, F. Zheng, L. Weng, Y. Ding, Overview of behavior recognition based on deep learning, *Artif. Intell. Rev.*, **2022** (2022), 1–33. https://doi.org/10.1007/s10462-022-10210-8
24. J. Liu, A. Shahroudy, D. Xu, A. C. Kot, G. Wang, Skeleton-based action recognition using spatio-temporal LSTM network with trust gates, in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, (2017), 3007–3021. https://doi.org/10.1109/TPAMI.2017.2771306
25. J. Liu, G. Wang, L. Y. Duan K. Abdiyeva, A. C. Kot, Skeleton-based human action recognition with global context-aware attention LSTM networks, in *IEEE Transactions on Image Processing*, IEEE, (2018), 1586–1599. https://doi.org/10.1109/TIP.2017.2785279
26. L. Wang, Z. Tong, B. Ji, G. Wu, TDN: Temporal difference networks for efficient action recognition, in 2*021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Nashville, USA, (2021), 1895–1904. https://doi.org/10.1109/CVPR46437.2021.00193
27. C. Liu, J. Ying, H. Yang, X. Hu, J. Liu, Improved human action recognition approach based on two-stream convolutional neural network model, *Vis. Comput.*, **37** (2021), 1327–1341. https://doi.org/10.1007/s00371-020-01868-8
28. C. Si, Y. Jing, W. Wang, L. Wang, T. Tan, Skeleton-based action recognition with spatial reasoning and temporal stack learning, in *Proceedings of the European Conference on Computer Vision*, ECCV, (2018), 103–118. https://doi.org/10.1007/978-3-030-01246-5_7

29. W. Yang, J. Zhang, J. Cai, Z. Xu, Shallow graph convolutional network for skeleton-based action recognition, *Sensors*, **21** (2021), 452. https://doi.org/10.3390/s21020452
30. Z. Chen, S. Li, B. Yang, Q. Li, H. Liu, Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition, in *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI, (2021), 1113–1122. https://doi.org/10.1609/aaai.v35i2.16197
31. C. Ding, S. Wen, W. Ding, K. Liu, E. Belyaev, Temporal segment graph convolutional networks for skeleton-based action recognition, *Eng. Appl. Artif. Intell.*, **110** (2022), 104675. https://doi.org/10.1016/j.engappai.2022.104675
32. L. Shi, Y. Zhang, J. Cheng, H. Lu, Two-stream adaptive graph convolutional networks for skeleton-based action recognition, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Long Beach, USA (2019), 12018–12027. https://doi.org/10.1109/CVPR.2019.01230
33. P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, N. Zheng, Semantics-guided neural networks for efficient skeleton-based human action recognition, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Seattle, USA, (2020), 1112–1121. https://doi.org/10.1109/CVPR42600.2020.00119
34. C. Si, W. Chen, W. Wang, L. Wang, T. Tan, An attention enhanced graph convolutional LSTM network for skeleton-based action recognition, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Long Beach, USA, (2019), 1227–1236. https://doi.org/10.1109/CVPR.2019.00132
35. S. Miao, Y. Hou, Z. Gao, M. Xu, W. Li, A central difference graph convolutional operator for skeleton-based action recognition, in *IEEE Transactions on Circuits and Systems for Video Technology,* IEEE, (2021), 4893–4899. https://doi.org/10.1109/TCSVT.2021.3124562
36. Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, W. Hu, Channel-wise topology refinement graph convolution for skeleton-based action recognition, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Montreal, Canada, (2021), 13359–13368. https://doi.org/10.1109/ICCV48922.2021.01311
37. T. Kipf, E. Fetaya, K. C. Wang, M. Welling, R. Zemel, Neural relational inference for interacting systems, in *International Conference on Machine Learning*, PMLR, (2018), 2688–2697.
38. A. Shahroudy, J. Liu, T. T. Ng, G. Wang, NTU RGB+D: A large scale dataset for 3D human activity analysis, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Las Vegas, USA, (2016), 1010–1019. https://doi.org/10.1109/CVPR.2016.115
39. W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, et al., The kinetics human action video dataset, preprint, arXiv:1705.06950.
40. Y. Du, W. Wang, L. Wang, Hierarchical recurrent neural network for skeleton based action recognition, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Boston, USA, (2015), 1110–1118. https://doi.org/10.1109/CVPR.2015.7298714
41. H. Liu, J. Tu, M. Liu, Two-stream 3D convolutional neural network for skeleton-based action recognition, preprint, arXiv:1705.08106.
42. H. H. Pham, H. Salmane, L. Khoudour, A. Crouzil, P. Zegers, S. A. Velastin, Spatio temporal image representation of 3D skeletal movements for view-invariant action recognition with deep convolutional neural networks, *Sensors*, **19** (2019), 1932. https://doi.org/10.3390/s19081932

43. Z. W. Huang, C. D. Wan, T. Probst, L. Van Gool, Deep learning on lie groups for skeleton-based ation recognition, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Honolulu, USA, (2017), 1243–1252. https://doi.org/10.1109/CVPR.2017.137
44. L. Bo, Y. Dai, X. Cheng, H. Chen, Y. Lin, M. He, Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN, in *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, IEEE, Hong Kong, (2017), 601–604. https://doi.org/10.1109/ICMEW.2017.8026282.
45. S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in *Thirty-second AAAI Conference on Artificial Intelligence*, AAAI, Palo Alto, USA, (2018), 7444–7452. https://doi.org/10.1609/aaai.v32i1.12328
46. C. Wu, X. J. Wu, J. Kittler, Spatial residual layer and dense connection block enhanced spatial temporal graph convolutional network for skeleton-based action recognition, in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, IEEE, Seoul, Korea (South), (2019), 1740–1748. https://doi.org/10.1109/ICCVW.2019.00216
47. Y. F. Song, Z. Zhang, C. Shan, L. Wang, Richly activated graph convolutional network for robust skeleton-based action recognition, in *IEEE Transactions on Circuits and Systems for Video Technology*, IEEE, (2021), 1915–1925. https://doi.org/10.1109/TCSVT.2020.3015051
48. H. Zhang, Y. Hou, P. Wang, Z. Guo, W. Li, Sar-nas: Skeleton-based action recognition via neural architecture searching, *J. Visual Commun. Image Represent.*, **73** (2020), 102942. https://doi.org/10.1016/j.jvcir.2020.102942
49. S. Cho, M. Maqbool, F. Liu, H. Foroosh, Self-attention network for skeletonbased human action recognition, in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, Snowmass, USA, (2020), 624–633. https://doi.org/10.1109/WACV45572.2020.9093639
50. C. Li, C. Xie, B. Zhang, J. Han, X. Zhen, J. Chen, Memory attention networks for skeleton-based action recognition, in *IEEE Transactions on Neural Networks and Learning Systems*, IEEE, (2021), 4800–4814. https://doi.org/10.1109/TNNLS.2021.3061115
51. B. Fernando, E Gavves, J Oramas, et al., Modeling video evolution for action recognition, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Boston, USA, (2015), 5378–5387. https://doi.org/10.1109/CVPR.2015.7299176