Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 4, pp. 2838–2850, DOI 10.12694/scpe.v25i4.2845

THE RECOGNITION OF AEROBICS MOVEMENTS USING BONE DATA COMBINED WITH ST-GCN

HUAHUA YANG[*], YANSHENG ZHAO[†], AND LI XIA [‡]

Abstract. To solve the aerobics action recognition and promote the gradual intelligence and standardization of aerobics teaching and evaluation, a network model based on spatial temporal graph convolution and combined with attention mechanism was proposed. This model improved the extraction efficiency of spatiotemporal features and channel features by introducing spatiotemporal graph and channel attention mechanisms, respectively, thereby improving the accuracy of action recognition. And a time extension module was introduced into each basic module, and additional features between adjacent vertices were extracted by extending the time graph between frames. These experiments confirmed that this model exhibited high accuracy in identifying aerobics movements. The recognition accuracy of basic actions was above 93.6%, and the recognition accuracy of two actions had reached 98.4% and 98.6% respectively. For advanced actions, the recognition accuracy of the model had slightly decreased, but the average value was still above 95%. The accuracy of difficult motion recognition had also achieved good results, reaching a maximum of 94.5%. These data indicate that this model can achieve high accuracy in handling action recognition tasks of different difficulty levels, and can identify aerobics movements of different difficulty levels.

Key words: Bone data; Graph convolution; Attention mechanism; Time extension; Multi-stream network

1. Introduction.

The evolution of computer vision and artificial intelligence has made action recognition increasingly important in various application scenarios. Aerobics, as a highly technical and artistic sports event, has extremely high requirements for the accuracy of movements [1-3]. The development of aerobics education has made great progress with the combination of computer learning and artificial intelligence technology. The movement recognition of aerobics provides effective help for accurate learning of aerobics and time love you. In addition, the motion recognition for aerobics has broad application prospects in education, sports rehabilitation, and personal fitness [4-5]. However, due to the complexity and diversity of aerobics movements, there are still many difficulties in using computers to achieve high-precision recognition of these movements. There is time dimension information in the aerobics movement, and the movement recognition needs to extract both spatial and temporal features. The lack of feature information will reduce the efficiency of the aerobics movement recognition. ST-GCN can extract spatial and temporal features at the same time, record joint motion information more completely, reduce the input of redundant information, and improve the efficiency of model recognition. Effective aerobics movement recognition can improve the irregular movement and improve the efficiency of teaching and training. Therefore, researching an efficient and accurate method for identifying aerobics movements has strong practical significance and scientific value. The research combines bone data and existing Spatial Temporal-Graph Convolutional Network (ST-GCN) to improve action recognition by introducing Attention Mechanism (AM), Time Extension Module (TEM), and multi-stream integration. Graph Attention Network(GAT) is introduced on the basis of ST-GCN to improve the poor flexibility of GCN. In order to strengthen the optimization of time graph, a time extension mode TEM is introduced creatively on the basis of this model. The model simply and effectively extracts the relevant features of multiple adjacent joints in human motion, which further improves the accuracy of the model. The aim is to achieve accurate recognition of aerobics actions and hope to provide valuable reference and inspiration for related fields, promoting intelligence

[*] Department of physical education, Guangzhou City University of Technology, Guangzhou, 510800, China (yangyuehua_2008@163.com)

[†] The school of Humanities and Social Sciences, Guangzhou Civil Aviation College, Guangzhou, 510800, China (zhaoyansheng@gcac.edu.cn)

[‡] College of Sports Training, Guangzhou Sports University, Guangzhou, 510500, China (13822273846@163.com)

and automation in aerobics teaching, training, and other aspects. The article first summarized the current research status of action recognition both domestically and internationally. Secondly, the improvement measures for ST-GCN were presented in detail. The third section described the results obtained from model validation through experiments. Finally, a comprehensive summary of this study was conducted, and the shortcomings and future prospects were proposed.

2. Related works. Early action recognition methods for bone data used manual feature extraction to model spatiotemporal correlations, without considering the internal connections between human joints. Deep learning effectively solved this problem. At present, there are three main methods in the human skeleton motion recognition: Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and GCN [6-7]. Muhammad K et al. proposed an AM based on Bidirectional Long-Short Term Memory (BiLSTM) to address the visual performance issues of existing motion recognition techniques in video frames during the training phase, and combined it with an expanded CNN. The recognition rates of this method on UCF11, UCF motion, and J-HMDB datasets reached 98.3%, 99.1%, and 80.2%, respectively [8]. Yang H et al. proposed a Space-Time Attention CNN that integrated temporal and spatial AM into a unified convolutional network, and mining time fragments through temporal attention mechanism. The non-moving regions are then focused through training by averaging the auxiliary classification losses of the pooling layer. The accuracy of the proposed model on UCF-101 dataset and HMDB-51 is 95.8% and 71.5%, respectively, and the performance is relatively advanced [9]. Zhang X et al. integrated spatial and temporal neighboring edges to represent an edge in a human skeleton graph, and designed a graph edge CNN. Experiments on the Kinetics and NTU-RGB+D datasets confirmed that the graph edge convolution effectively captures moving features, and CNN's graph edge was superior to other advanced existing bone-based action recognition methods [10].

In RNN-based methods, human skeleton data are treated as a series of vectors, which help to better explore the temporal correlations between human skeleton data. Avola D et al. proposed a novel method for motion recognition in video sequences by combining 2D bone data and dual branch stacked LSTM. This method used 2D bones reconstructed from RGB video streams and used 3D-CNN to process missing bone data. The comparison on the UT Kinect and NTU-RGB+D datasets confirmed that the accuracy of this method was completely equivalent to that of works based on 3D bones [11]. Xu S et al. integrated GCN into LSTM, utilized skeletal body structure information and enhanced multi-level co-occurrence feature learning, and used parallel LSTM to model the temporal dynamics of aggregated features for action recognition. The proposed model has obvious superiority on NTU RGB+D 60/120 data set and NTU RGB+D 60/120 data set [12]. Zhu A et al. proposed an end-to-end bidirectional LSTM-CNN and employed a layered spatiotemporal dependency model to explore the rich spatiotemporal information in bone data. In the fusion framework of CNN and LSTM, bone data were constructed from a dependency model and used as input to the proposed network. Its effectiveness had been verified on different datasets [13].

Human skeleton data exists in the form of graphs, and using topology maps is more suitable for expressing skeleton data. Shi L et al. proposed a novel multi-stream attention enhancement adaptive GCN. This structure could be learned in an end-to-end manner based on input data. In addition, a multi-stream framework including moving data had shown significant improvements in recognition accuracy [14]. Zhang Z et al. proposed a novel structure feature fusion adaptive GCN model, which could effectively fuse the topological structure of bone maps and joint features. Experiments on different datasets had confirmed that the improved GCN outperformed existing state-of-the-art methods, with an average accuracy improvement of over 0.6% [15]. Tsai M F and Chen C H proposed a new training way for emotion recognizing, which used bone detecting way to obtain bone points' changing degree and used nearest neighbor algorithms for classifying speed. Compared with the basic method, the ST-GCN technology that considered changes could enhance recognizing accuracy by over 50% [16].

Based on relevant research both domestically and internationally, GCN has become a hot topic for human skeleton motion recognition due to its excellent performance in graph data with topological structures. Among them, ST-GCN is a groundbreaking research work that successfully applies GCN to human skeleton motion recognition tasks. However, ST-GCN still has drawbacks such as inflexible weight allocation and failure to distinguish redundant information. Therefore, based on bone data and combined with GCN, the study aims to use computer technology to achieve higher precision action recognition and apply it to the intelligent teaching of aerobics.

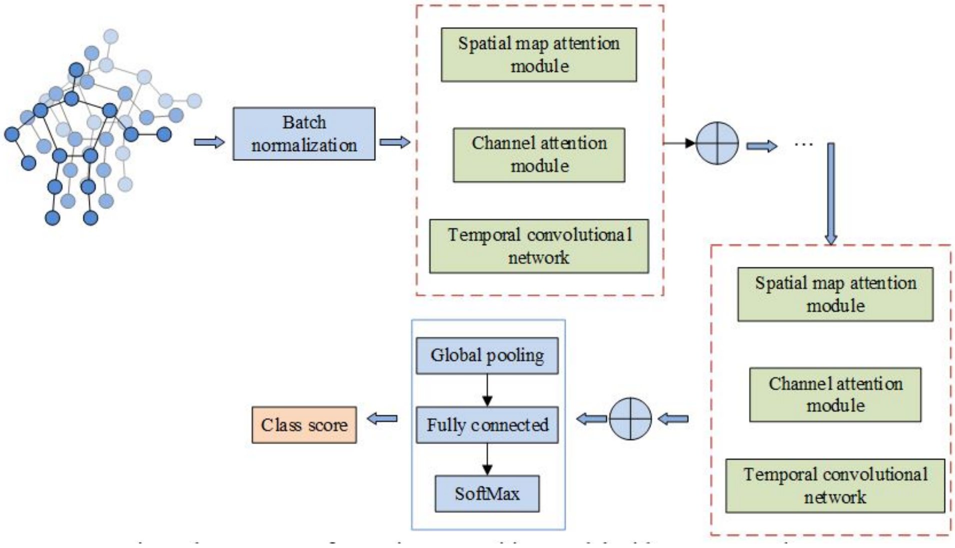


Fig. 3.1: The structure of an action recognition model with ST-GCN and AM

3. The recognition and improvement of aerobics movements using bone data combined with ST-GCN. On the basis of analyzing the existing defects of ST-GCN and combining the characteristics of bone data, an improved action recognition model is proposed. Firstly, for the weight allocation in GCN, GAT is used to improve the original GCN, and a Channel Attention Module (CAM) is added to focus on more important channel information. Although GAT can extract spatiotemporal features, it is easy to overlook the optimization of inter frame time maps. Therefore, the study further introduces TEM and proposes an integrated framework using multi-stream networks to fuse information to improve the accuracy of action recognition.

3.1. An action recognition model based on ST-GCN combined with AM. The composition of human bones and joints can naturally form a topological map. GCN has excellent performance in processing graph structured data, therefore it is widely used in action recognition based on human skeleton. However, for human motion data, the human motion skeleton map is composed of time series, and the time information covers the spatial characteristics at each time node and the temporal characteristics between frames [17]. Therefore, how to cleverly use graph convolution to extract spatiotemporal features in human motion has become a research focus of skeleton-based action recognizing way. A spatiotemporal graph attention model based on ST-GCN combined with AM is proposed to improve GCN'S poor flexibility and action recognizing accuracy in Figure 3.1.

From Figure 3.1, the improved spatiotemporal graph attention model constructs a human skeleton spatiotemporal graph using human skeleton data, and then introduces GAT [18] and multi-head graph AM on the basis of ST-GCN to enhance model's performance. In addition, to further improve its accuracy, CAM is introduced to make the model concerned with more important channel information. The basic unit of the Spatiotemporal Graph Attention Model (SGAM) consists of three parts: firstly, the spatial graph attention module, secondly, CAM, and finally, the time convolution module with a kernel size of 9. The spatial graph attention module utilizes Graph Attention Network (GAT) to partially replace the original GCN to improve recognition rate. The proposed spatial graph attention module includes a GAT layer that can better extract spatiotemporal layer features, as well as a set of batch regularization layers and a set of ReLU activation function layers. Then there is CAM, which uses AM for channels with different semantic features, making the network more focused on more important channel characteristics, thereby reducing redundant features. And each unit contains a residual module. ST-GCN applies graph convolution to the spatiotemporal graph structure, and Figure 2 shows the overall processing flow.

From Figure 3.1, ST-GCN first extracts the skeleton sequence of actions from the input video using a

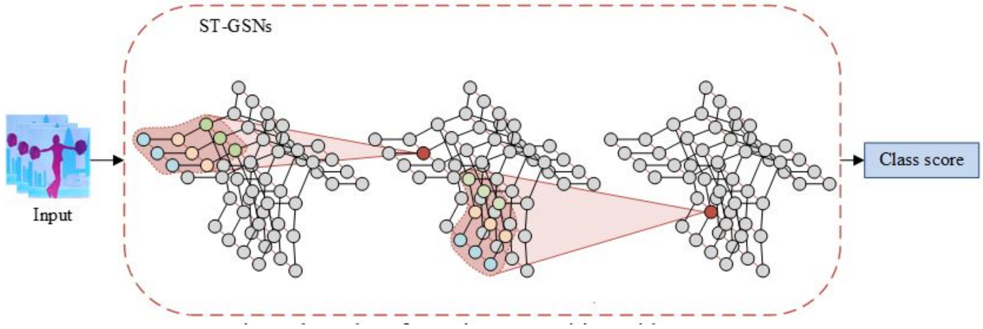
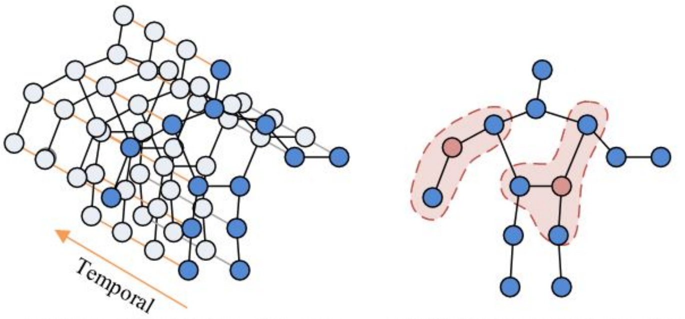


Fig. 3.2: Flow chart for action recognition with ST-GCN



(a) Time-space diagram of human skeleton

(b) Neighborhood sampling area with a distance of 1

Fig. 3.3: The spatiotemporal structure of human skeleton and its sampling function

pose estimation algorithm, and constructs a spatiotemporal map of the human skeleton based on the skeleton sequence. Then, the joint coordinates of the human body are interconnected as input to ST-GCN, and multiple graph convolution modules are used to automatically extract spatiotemporal features [19]. Each module consists of GCN and TCN. Finally, the results of action classification are obtained through a SoftMax classifier. When constructing a spatiotemporal map of human skeleton, its spatiotemporal sequence is connected into a spatiotemporal map $G = (V, E)$ according to the given rules. V represents the set of joint points in a spatiotemporal graph. E represents a set of edges. The set of joint points includes the natural joint points of the human body on each frame, denoted as $V = \{v_{ti} \,|t = 1, 2, \cdots, T, i = 1, 2, \cdots, N\}$. T is the amount of frames in a continuous video. N is the amount of human joint points at a certain frame. The set of edges is composed of two subsets: spatial edge $E_S = \{v_{ti}v_{tj} \,|(i, j) \in H\}$ and temporal edge $E_F = \left\{v_{ti}v_{(t+1)i}\right\}$. H is a set of naturally connected human joints. In the spatial graph attention module, formula (3.1) is the sampling function $p\left(v_{tj}, v_{ti}\right)$ of the graph structure.

$$\left\{ \begin{array}{l} B\left(v_{tj}\right) = \{v_{tj} \,|d\left(v_{tj}, v_{ti}\right) \leq D\} \\ p\left(v_{tj}, v_{ti}\right) = v_{tj} \end{array} \right. \tag{3.1}$$

In formula (3.1), $B\left(v_{tj}\right)$ is the set of neighboring pixels in the graph. $d\left(v_{tj}, v_{ti}\right)$ is the shortest distance between v_{tj} and v_{ti}. D is the distance threshold. Figure 3 shows human skeleton's spatiotemporal structure and the neighborhood sampling area with a distance of 1 from the human skeleton.

To obtain the attention coefficient of the spatial graph, the feature dimension of the central node and its surrounding neighboring nodes is first improved. For the central node v_{ti} and its neighboring node v_{tj} within

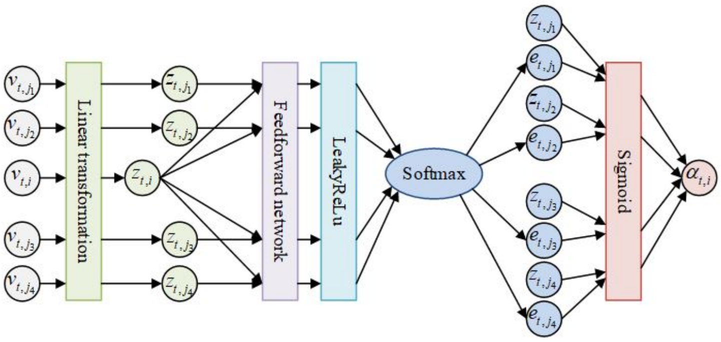


Fig. 3.4: The calculation of attention coefficients in spatial maps

a certain frame t, a weight matrix W is introduced to achieve shared linear transformation of the nodes, to obtain higher dimensional features. Formula (3.2) is a node linear transformation.

$$\begin{cases} z_{ti} = W f_{in}(v_{ti}) \\ z_{tj} = W f_{in}(v_{tj}) \end{cases} \tag{3.2}$$

In formula (3.2), $f_{in}(v_{ti}) \in \mathrm{R}^C$ and $f_{in}(v_{tj}) \in \mathrm{R}^C$ represent the input feature vectors of two nodes, respectively. The transformed vectors z_{ti} and z_{tj} are concatenated to obtain a new dimension vector, and the attention value e_{tij} of the center node i on node j is calculated in frame t in formula (3.3).

$$e_{tij} = LeakR\left(\vec{a}^{\mathrm{T}}\left[z_{ti} \| z_{tj}\right]\right), j \in B(v_{ti}) \tag{3.3}$$

In formula (3.3), $LeakR(\bullet)$ is the activation function. a represents a single-layer feedforward neural network. Formula (3.4) is the attention coefficient α_{tij} of each neighboring node towards the central node v_{ti}.

$$\alpha_{tij} = \frac{\exp\left(LeakR\left(\vec{a}^{\mathrm{T}}\left[z_{ti} \| z_{tj}\right]\right)\right)}{\sum_{k\in B(v_{ti})} \exp\left(LeakR\left(\vec{a}^{\mathrm{T}}\left[z_{ti} \| z_{tj}\right]\right)\right)} \tag{3.4}$$

Figure 3.4 shows the calculation of attention coefficient in spatial maps.

According to the attention coefficient, the final output $f_{out}(v_{ti})$ of the central node in the single head graph AM can be obtained in formula (3.5).

$$\left(\sum_{j\in B(v_{ti})} \alpha_{tij} W f_{in}\left(p\left(v_{tj}, v_{ti}\right)\right)\right) = \sigma\left(\sum_{j\in B(v_{ti})} a_{tij} z_{tj}\right) \tag{3.5}$$

In formula (3.5), σ is the activation function. The same weight when calculating the output of a single graph's attention is not conducive to the learning ability. Therefore, this study used multi-head spatial graph attention to assign different attention coefficients and weights to different features. Then, the calculated features of $f_{out}(v_{ti}) = \Big\|_{k=1}^{K} \sigma\left(\sum_{j\in B(v_{ti})} a_{tij}^{k} z_{tj}^{k}\right)$ independent AMs are concatenated to obtain the final output of each node in formula (3.6).

$$f_{out}(v_{ti}) = \Bigg\|_{k=1}^{K} \sigma\left(\sum_{j\in B(v_{ti})} a_{tij}^{k} z_{tj}^{k}\right) \tag{3.6}$$

By extending the model from the spatial domain to the temporal domain through a dataset composed of different frames of the same joint point, the final set of neighborhood nodes $B^*(v_{ti})$ for v_{ti} is obtained in formula (3.7).

$$B^*(v_{ti}) \to \left\{v_{qj} \left| d(v_{tj}, v_{ti}) \le K, |q-t| \le \lfloor \Gamma/2 \rfloor \right.\right\} \tag{3.7}$$

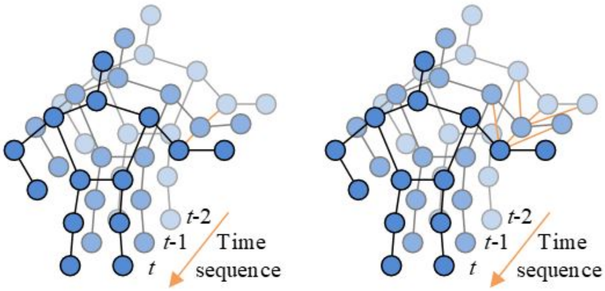


Fig. 3.5: Comparison of spatiotemporal graph connection methods

In formula (3.7), Γ controls the size of the convolution kernel in the time domain. CAM is introduced after the attention module of the spatial map, and the final output $f^{*}_{out}\left(v_{ti}\right)$ is obtained in formula (3.8).

$$f^{*}_{out}\left(v_{ti}\right)=f_{out}\left(v_{ti}\right)\cdot s_{c} \tag{3.8}$$

3.2. ST-GCN-GAT action recognition based on multi-stream extension improvement. GAT can extract spatiotemporal features, but it focuses more on representing a more suitable spatial map within frames, which can easily overlook the optimization of inter frame time maps. Therefore, the study introduces a TEM into the model, which expands the time graph by adding connections to adjacent multiple vertices based on inter frame and additional feature extraction. In previous recognition models, most only used joint point data of the human skeleton. Aerobics movements are continuous movements that combine multiple parts. Therefore, to fully utilize bone data, enhance the expression ability and recognition accuracy of the model, a multi-stream network integrated framework model is proposed, in which each network uses the same structure to fuse multiple types of information. Ordinary time convolution can only extract the information of neighboring nodes in the same frame, and cannot extract the information of the same related node and its neighboring nodes between frames. Figure 3.5 shows the connections of the traditional spatiotemporal map [20] and the improved TEM spatiotemporal map.

In Figure 3.5 (b), TEM adds edges to vertices corresponding to the same joint and adds edges to vertices corresponding to multiple adjacent joints between frames, and calculates convolution based on the same multiple vertices between frames. TEM is used to extract inter frame features, as it is added between the spatial attention layer and temporal convolution to expand temporal dimension's sampling area. For the temporal convolution outputting of TEM, formula (3.9) is the $f^{*}_{out}\left(v_{ti}\right)$ of the i-th vertex v_i in space at time t.

$$f_{out}\left(v_{ti}\right)=\sigma\left(\sum_{v_{(t-1)j}\in B^{T}\left(v_{ti}\right)}\alpha_{i,j}Wf_{in}\left(v_{t-1,j}\right)\right) \tag{3.9}$$

In formula (3.9), $\alpha_{i,j}$ represents the attention coefficient between nodes j and i. $B^{T}\left(v_{ti}\right)$ represents the inter frame sampling area of v_{ti} in formula (3.10).

$$B^{T}\left(v_{ti}\right)=\left\{v_{(t-1)j}\left|d\left(v_{(t-1)j},v_{(t-1)i}\right)\leq D^{T}\right.\right\} \tag{3.10}$$

In formula (3.10), D^{T} is the maximum length of inter frame sampling. After introducing the SGAM and CAM mechanisms, TEM is introduced to expand time dimension's sampling area and extract the temporal changes of multiple joints in motion. However, the entire network does not fully utilize the skeleton time sequence diagram constructed from human bone data. Therefore, the study proposes to use the same structure to process all data streams in a multi-stream structure, integrating joint and bone information along with motion information into one framework. Figure 6 shows a multi-stream network.

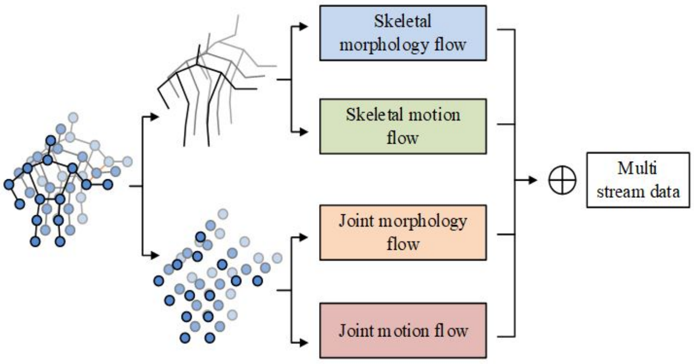


Fig. 3.6: The structure of multi-stream networks

From Figure 3.5, there are four types of branches in the multi-stream network, which are inputted into four streams. The final result is obtained by adding four streams' SoftMax scores. Each human skeleton has a center of gravity point, so the source joint point $v_{i,t} = (x_{i,t}, y_{i,t}, z_{i,t})$ in a pair of adjacent joint points is defined as the node closer to the skeleton's center of gravity, and the other node is defined as the target node $v_{j,t} = (x_{j,t}, y_{j,t}, z_{j,t})$. The vector from the source joint point to the target joint point can represent the length and direction of the bone between two joint points, i.e., the vector $b_{i,j,t}$ of bone in formula (3.11).

$$b_{i,j,t} = (x_{j,t} - x_{i,t}, y_{j,t} - y_{i,t}, z_{j,t} - z_{i,t}) \tag{3.11}$$

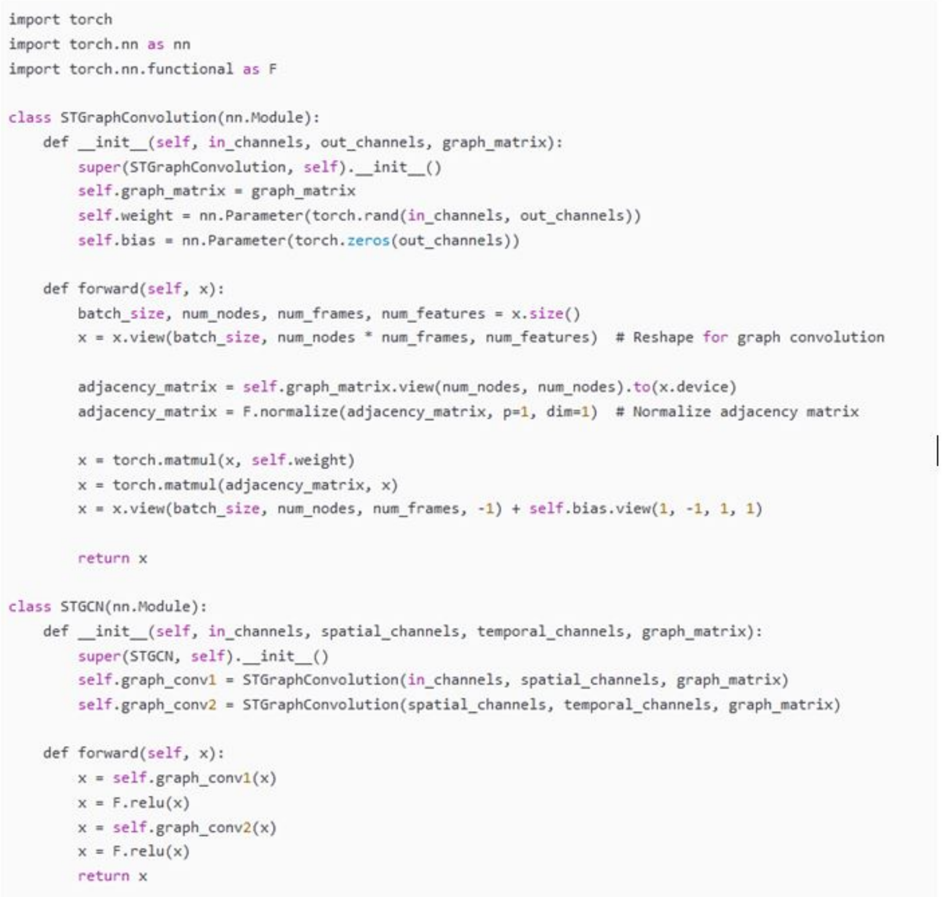
Because the skeleton graph is an acyclic data structure, there will eventually be an extra root node without assigning bones. Assigning an empty bone to the root node can design the bone and joint into the same graph and network. Previous studies have confirmed that optical flow fields are suitable for human motion recognition based on RGB videos. Therefore, the study defines the joint coordinate difference between consecutive frames as joint motion information, while the bone coordinate difference is defined as bone motion information. Given the connection $v_{i,t} = (x_{i,t}, y_{i,t}, z_{i,t})$ of frame t and the same joint point $v_{i,t+1} = (x_{i,t+1}, y_{i,t+1}, z_{i,t+1})$ in frame $t + 1$, then formula (3.12) shows the motion information $m_{i,t,t+1}$ between them.

$$m_{i,t,t+1} = (x_{i,t+1} - x_{i,t}, y_{i,t+1} - y_{i,t}, z_{i,t+1} - z_{i,t}) \tag{3.12}$$

The basic Pseudo code for the model is shown in the figure 3.7.

4. Performance analysis of aerobics movement recognition using bone data combined with ST-GCN. To verify the action recognition ability of this proposed model, multiple tests were conducted on two public datasets and a self-made aerobics action dataset. They included action recognition effect verification based on ST-GCN combined with AM, TEM effect verification, and multi-stream network effect verification.

4.1. Action recognition effect based on ST-GCN combined with AM. Two common datasets used in this study were NTU-RGB+D and Kinetic. NTU-RGB+D was collected by the Kinect V2 depth sensor using three different camera angles, containing a total of 56880 video samples for 60 types of actions. The standards for dividing training and testing sets using NTU-RGB+D included Cross-Subject and Cross-View. These actions were performed by 40 people, and a total of 25 skeleton points were collected in the datasets. Kinetics contained the actions of 400 categories of characters, each with at least 400 video clips, each lasting approximately 10 seconds. In order to obtain bone data from Kinetics, the captured raw video is cropped and the frame rate converted. A demonstration by a professional aerobics' teacher was used to shoot and produce sample videos, completing the aerobics dataset. The self-made dataset contained 1800 sample videos, divided into a training set and a testing set in an 8:2 ratio. The spatial dimension features started from 3, and the output feature dimensions for the first three layers, middle three layers, and last three layers were 64, 128,



```
import torch
import torch.nn as nn
import torch.nn.functional as F

class STGraphConvolution(nn.Module):
    def __init__(self, in_channels, out_channels, graph_matrix):
        super(STGraphConvolution, self).__init__()
        self.graph_matrix = graph_matrix
        self.weight = nn.Parameter(torch.rand(in_channels, out_channels))
        self.bias = nn.Parameter(torch.zeros(out_channels))

    def forward(self, x):
        batch_size, num_nodes, num_frames, num_features = x.size()
        x = x.view(batch_size, num_nodes * num_frames, num_features)  # Reshape for graph convolution

        adjacency_matrix = self.graph_matrix.view(num_nodes, num_nodes).to(x.device)
        adjacency_matrix = F.normalize(adjacency_matrix, p=1, dim=1)  # Normalize adjacency matrix

        x = torch.matmul(x, self.weight)
        x = torch.matmul(adjacency_matrix, x)
        x = x.view(batch_size, num_nodes, num_frames, -1) + self.bias.view(1, -1, 1, 1)

        return x

class STGCN(nn.Module):
    def __init__(self, in_channels, spatial_channels, temporal_channels, graph_matrix):
        super(STGCN, self).__init__()
        self.graph_conv1 = STGraphConvolution(in_channels, spatial_channels, graph_matrix)
        self.graph_conv2 = STGraphConvolution(spatial_channels, temporal_channels, graph_matrix)

    def forward(self, x):
        x = self.graph_conv1(x)
        x = F.relu(x)
        x = self.graph_conv2(x)
        x = F.relu(x)
        return x
```

Fig. 3.7: The basic Pseudo code of the model

Table 4.1: Laboratory hardware and software environment setup

Hardware and software configuration	Version model
CPU	Intel(R)Core i7-9700K
GPU	GTX 1060
Operating system	Ubuntu 16.04.6
RAM	16G
Display memory	6G
Development editor	Pycharm
frameworks	Pycaffe
Python version	2.7

and 256, respectively. After processing through basic units, dropout features were randomly selected with a probability of 0.5. The learning rate was 0.01, multiplied by 0.1 for every 10 epochs, and the batchsize was set to 64. Laboratory environment Settings are shown in Table 1.

The value range of K in multi-head attention was $[2, 9]$. ST-GCN was combined with GAT to construct a recognition model, and the influence of K in multi-head attention was verified on NTU-RGB+D. Figure 8 showed the top-1 results obtained.

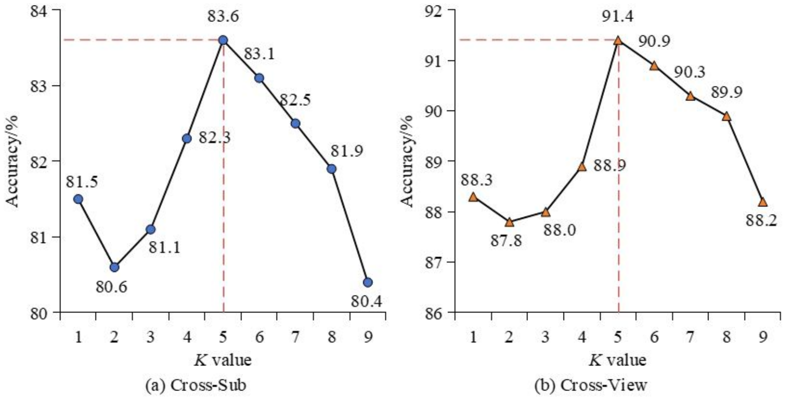


Fig. 4.1: The effect of K on the validation of multiple head attention in NTU-RGB+D

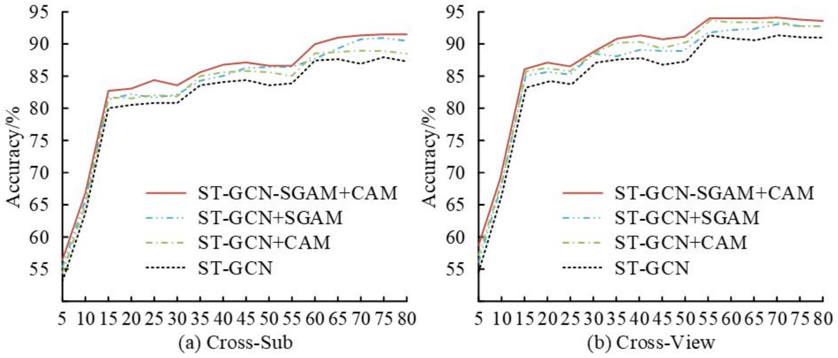
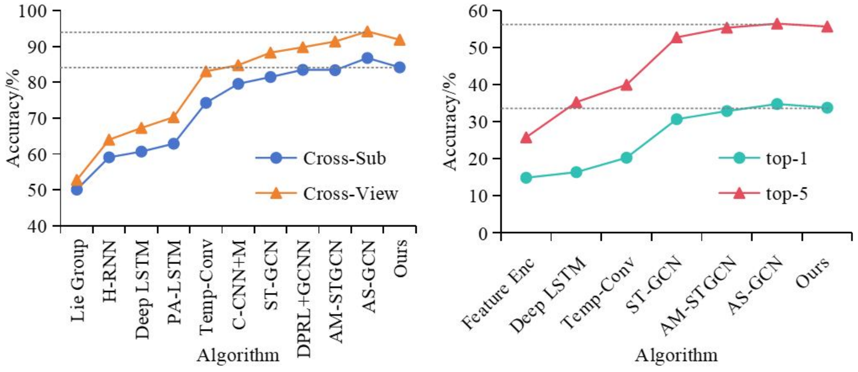


Fig. 4.2: Training curves of different models on datasets

In Figure 4.1, for multi-head AM, the accuracy of this model gradually increased with the increase of K. But before $K = 4$, the performance of this model was not as good as the benchmark ST-GCN. From Figure 8 (a), after K equaled 2, the accuracy first gradually improved, then reached the highest recognition rate of 83.6% when K equaled 5, and then gradually decreased. This indicated that selecting an appropriate K in multi-head attention might improve this recognition model's accuracy, which was consistent with the results obtained by dividing the training set according to Cross-View standard. Therefore, $K = 5$ was set in subsequent experiments. Figure 4.2 showed the training curves of the model in NTU-RGB+D and Kinect.

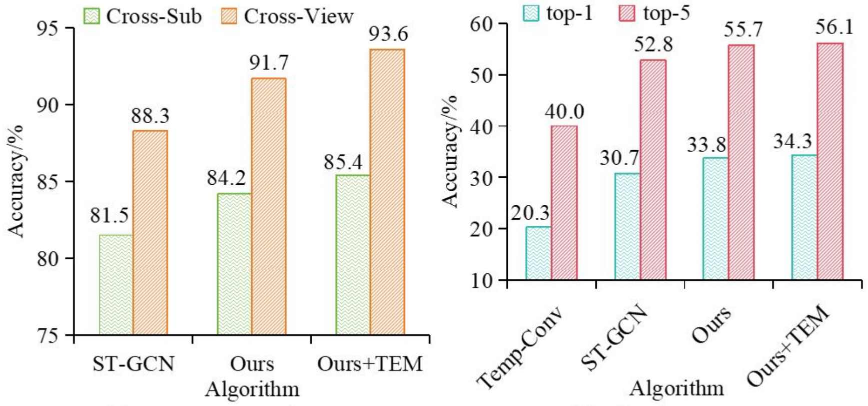
From Figure 4.2, although all four models tended to stabilize around 60 rounds, the accuracy of ST-GCN with the introduction of SGAM and CAM was higher than that of other models in 80 rounds of training. The study selected Lie Group, Feature Encoding (Feature Enc), Hierarchical-RNN (H-RNN) based method, Deep LSTM, Part-Aware LSTM (PA-LSTM), Temporal Convolution (Temp-Conv), Clip CNN + Muti-task learning (C-CNN+M), ST-GCN, Deep Progressive Reinforcement Learning + GCNN (DPRL+GCNN), and AM-STGCN, Actional-Structural GCN (AS-GCN) to conduct comparative experiments in Figure 4.3.

From Figure 4.3 (a), models such as Temp-Conv, ST-GCN, DPRL+GCNN achieved scores of over 80% on NTU-RGB+D, especially AS-GCN, which achieved scores of 86.8% and 94.2% on Cross-Sub and Cross-View indicators. ST-GCN, which introduced SGAM and CAM, also performed quite well on these two indicators, with scores of 84.2% and 91.9%. Although it was slightly lower than AS-GCN, the difference was not significant. These experiments confirmed that although the introduction of attention modules could effectively improve



(a) NTU-RGB+D dataset (b) Kinetic Dataset

Fig. 4.3: Comparative experimental results on two datasets



(a) NTU-RGB+D dataset (b) Kinetic Dataset

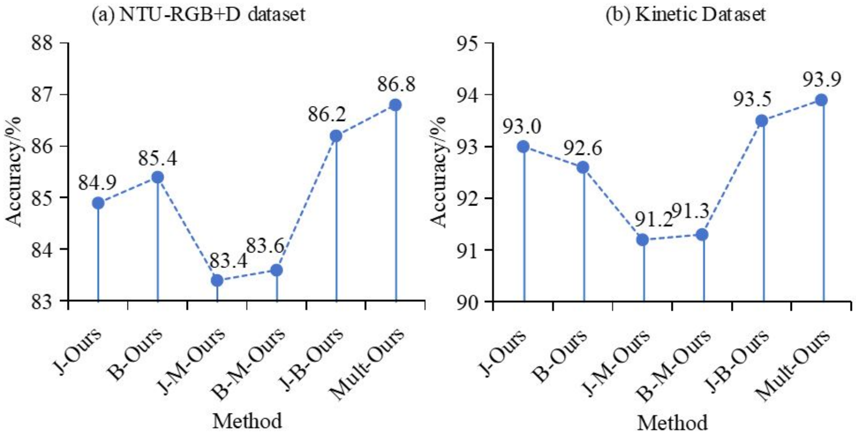
Fig. 4.4: Verification of the optimization effect of TEM

model performance, focusing solely on the relationship between physical adjacent joint points on the skeleton graph was flawed.

4.2. ST-GCN-GAT action recognition effect based on multi-stream extension improvement. The configuration and evaluation indicators of the optimization validation experiment for TEM and multi-stream networks were the same as before. To verify the improvement effect of TEM on the model, ablation experiments were conducted on two datasets in Figure 11. ”Ours” in Figure 4.4 represented ST-GCN+SGAM+CAM.

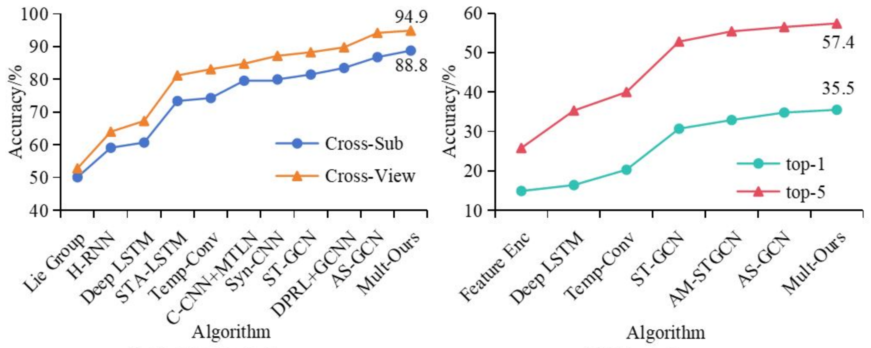
From Figure 4.4 (a), after introducing TEM, the recognition accuracy in Cross-Sub and Cross-View forms had been improved to 85.4% and 93.6%, respectively. In Figure 4.4 (b), before introducing TEM, the top-1 and top-5 accuracy of ST-GCN+SGAM+CAM were 33.8% and 55.7%, respectively. However, after introducing TEM, the scores of these indicators increased to 34.3% and 56.1%. Although the improvement was relatively small, it still indicated that TEM had a certain optimization effect on this model. Figure 4.5 showed a comparative experiment of multi-stream networks on NTU-RGB+D. J represented joint morphology, B represented bone morphology, J-M represented joint motion flow morphology, and B-M represented bone motion flow morphology. The ”Ours” here represents ST-GCN+SGAM+CAM+TEM.

From Figure 4.5, the accuracy of the model that only showed joint and bone morphology was similar, with 84.9%, 85.4%, and 93.0%, 92.6%, respectively. Considering the morphology of joints and bones simultaneously,



(a) Cross-Sub (b) Cross-View

Fig. 4.5: Results of multiteam network comparison experiment on NTU-RGB+D



(a) NTU-RGB+D dataset (b) Kinetic Dataset

Fig. 4.6: Comparative experimental results of optimized models

the model's effectiveness would be improved, with accuracy rates of 86.2% and 93.5%. After the introduction of multi-stream networks, the accuracy of this model was further improved, reaching 86.8% and 93.9%. This indicated that using a multi-stream network would help improve the model performance. Then, the optimized model was compared with several more advanced methods on NTU-RGB+D and Kinect in Figure 4.6.

From Figure 4.6 (a), the model further optimized using TEM and multi-stream network achieved recognition accuracy of 88.8% and 94.9% under Cross-Sub and Cross-View standards, respectively, achieving better performance than AS-GCN. In Figure 4.6 (b), in Kinect, the improved model had the highest accuracy of top-1 and top-5 among all models, with 35.5% and 57.4%, respectively. This indicated that the optimization effect of this model was very significant, and the optimization strategy could effectively improve the accuracy of model prediction. Multiple models are verified on self-built data sets, and the results obtained are shown in Table 4.2.

As shown in Table 4.2, the model proposed in this study has absolute advantages in accuracy, accuracy, recall and F1-score indicators. The recall rate and F1-score of the model proposed in this study were 89.9% and 91.0%, respectively, with high aerobics action recognition ability. H-RNN model, on the other hand, has poor performance and can not accurately identify aerobics movements. Finally, comparative experiments were conducted on a self-made aerobics dataset to verify the model effectiveness in identifying aerobics movements

Table 4.2: The results of index validation of multiple models

Model	Accuracy%	Precision%	Recall rate%	F1-score%
Mult-Ours	91.8	94.8	89.9	91.0
AS-GCN	90.6	92.5	89.1	90.5
AM-STGCN	86.7	89.0	84.7	89.4
H-RNN	82.1	86.4	81.2	86.6
STA-LSTM	85.4	88.7	83.5	88.9

Table 4.3: The recognition results on a self-made aerobics dataset

Action classification	Action number	Accuracy/%
Basic Actions	A1	97.3
	A2	95.7
	A3	93.6
	A4	98.4
	A5	98.6
Advanced Actions	B1	95.4
	B2	97.6
	B3	93.2
	B4	94.5
	B5	96.2
	B6	95.7
Difficult actions	C1	93.2
	C2	90.3
	C3	91.4
	C4	92.8
	C5	94.5

in Table 4.3.

In Table 4.3, the improved model had good recognition performance for various movements in the aerobics dataset, with an accuracy rate of over 90%. For the basic action part, the recognition accuracy of all five actions was above 93.6%, among which the A4 and A5 actions had the highest recognition accuracy of 98.4% and 98.6%, respectively. For advanced actions, the recognition accuracy of this model had slightly decreased, but the average value was still above 95%. For Difficult actions, although the average recognition accuracy had slightly decreased, the recognition accuracy of C5 actions still reached 94.5%, and the lowest C2 had an accuracy of 90.3%. On the whole, the recognition accuracy of basic movements is high, and the average recognition accuracy decreases gradually with the increase of movement difficulty. Difficult actions often have multiple node data and change rapidly in a short time, and the model can not quickly extract all of them. These data indicated that when identifying difficult aerobics movements, the recognition accuracy of the model could still be maintained at a high level.

5. Conclusion. To achieve accurate recognition of aerobics movements, an improved motion recognition model was proposed by combining bone data and ST-GCN. This model enhanced the extraction of spatiotemporal and channel features by utilizing SGAM and CAM mechanisms, thereby improving action recognizing accuracy. In response to the optimization of inter frame time maps and the sufficient utilization of skeleton information, a TEM was introduced into each basic module, and a multi-stream network containing node information, skeleton information, and their motion information was proposed. This enabled the model to extract information from neighboring nodes between frames and extended the time map on the basis of additional feature extraction. The testing on the self-made aerobics' dataset confirmed that the recognition accuracy of the model for basic movements was above 93.6%. Even in the face of difficult action recognition, the accuracy of this model was excellent, such as the recognition accuracy of C5 actions still reaching 94.5%, with the lowest C2 being 90.3%. These experiments confirmed that the model had achieved high accuracy in processing basic, advanced, and difficult motion recognition, which helped to promote the development and application of aerobics motion recognition. However, the study only considered motion recognition in single person aerobics. Further optimization can be made for recognition in interactive scenarios between humans and objects. The average accuracy of the model has decreased when it recognizes difficult movements. The subsequent research can improve the average accuracy of the model for difficult movements of aerobics. In addition, the developed model can identify aerobics movements more accurately, and subsequent work can generalize the model and apply it to more scenarios, such as the specific challenges of aerobic exercise.

REFERENCES

[1] Yan, G. & Woźniak, M. Accurate key frame extraction algorithm of video action for aerobics online teaching. *Mobile Networks And Applications.* **27**, 1252-1261 (2022)

[2] Fuxiang, L. Adaptive recognition method of aerobics decomposition action image based on feature extraction. *Science Technology And Engineering.* **476**, 153-158 (2019)

[3] Yan, G. & Woźniak, M. Accurate key frame extraction algorithm of video action for aerobics online teaching. *Mobile Networks And Applications.* **27**, 1252-1261 (2022)

[4] Li, L. An online arrangement method of difficult actions in competitive aerobics based on multimedia technology. *Security And Communication Networks.* **2021** pp. 1-12 (2021)

[5] Yao, G., Lei, T. & Zhong, J. A review of convolutional-neural-network-based action recognition. *Pattern Recognition Letters.* **118** pp. 14-22 (2019)

[6] Ahmad, T., Jin, L., Zhang, X., Lai, S., Tang, G. & Lin, L. Graph convolutional neural network for human action recognition: A comprehensive survey. *IEEE Transactions On Artificial Intelligence.* **2**, 128-145 (2021)

[7] Zhang, C., Liang, J., Li, X., Xia, Y., Di, L., Hou, Z. & Huan, Z. Human action recognition based on enhanced data guidance and key node spatial temporal graph convolution. *Multimedia Tools And Applications.* **81**, 8349-8366 (2022)

[8] Muhammad, K., Ullah, A., Imran, A., Sajjad, M., Kiran, M., Sannino, G. & Albuquerque, V. Human action recognition using attention-based LSTM network with dilated CNN features. *Future Generation Computer Systems.* **125** pp. 820-830 (2021)

[9] Yang, H., Yuan, C., Zhang, L., Sun, Y., Hu, W. & Maybank, S. STA-CNN: Convolutional spatial-temporal attention learning for action recognition. *IEEE Transactions On Image Processing.* **29** pp. 5783-5793 (2020)

[10] Zhang, X., Xu, C., Tian, X. & Tao, D. Graph edge convolutional neural networks for skeleton-based action recognition. *IEEE Transactions On Neural Networks And Learning Systems.* **31**, 3047-3060 (2019)

[11] Avola, D., Cascio, M., Cinque, L., Foresti, G., Massaroni, C. & Rodolà, E. 2-D skeleton-based action recognition via two-branch stacked LSTM-RNNs. *IEEE Transactions On Multimedia.* **22**, 2481-2496 (2019)

[12] Xu, S., Rao, H., Peng, H., Jiang, X., Guo, Y., Hu, X. & Hu, B. Attention-based multilevel co-occurrence graph convolutional LSTM for 3-D action recognition. *IEEE Internet Of Things Journal.* **8**, 15990-16001 (2020)

[13] Zhu, A., Wu, Q., Cui, R., Wang, T., Hang, W., Hua, G. & Snoussi, H. Exploring a rich spatial–temporal dependent relational model for skeleton-based action recognition by bidirectional LSTM-CNN. *Neurocomputing.* **414** pp. 90-100 (2020)

[14] Shi, L., Zhang, Y., Cheng, J. & Lu, H. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *IEEE Transactions On Image Processing.* **29** pp. 9532-9545 (2020)

[15] Zhang, Z., Wang, Z., Zhuang, S. & Huang, F. Structure-feature fusion adaptive graph convolutional networks for skeleton-based action recognition. *IEEE Access.* **8** pp. 228108-228117 (2020)

[16] Tsai, M. & Chen, C. Spatial temporal variation graph convolutional networks (STV-GCN) for skeleton-based emotional action recognition. *IEEE Access.* **9** pp. 13870-13877 (2021)

[17] Peng, W., Shi, J., Varanka, T. & Zhao, G. Rethinking the ST-GCNs for 3D skeleton-based human action recognition. *Neurocomputing.* **454** pp. 45-53 (2021)

[18] Shen, N., Feng, Z., Li, J., You, H. & Xia, C. Action fusion recognition model based on GAT-GRU binary classification networks for human-robot collaborative assembly. *Multimedia Tools And Applications.* **82**, 18867-18885 (2023)

[19] Wang, X., Cheng, M., Eaton, J. & Heieh, C. Fake node attacks on graph convolutional networks. *Journal Of Computational And Cognitive Engineering.* **1**, 165-173 (2022)

[20] Fang, Y., Luo, B., Zhao, T., He, D., Jaing, B. & Liu, Q. ST-SIGMA: Spatio-temporal semantics and interaction graph aggregation for multi-agent perception and trajectory forecasting. *CAAI Transactions On Intelligence Technology.* **7**, 744-757 (2022)

Edited by: Zhengyi Chai
Special issue on: Data-Driven Optimization Algorithms for Sustainable and Smart City
Received: Nov 14, 2023
Accepted: May 6, 2024