Computer Modeling in Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2023.027500

ARTICLE

Building Indoor Dangerous Behavior Recognition Based on LSTM-GCN with Attention Mechanism

Qingyue Zhao[1], Qiaoyu Gu[2], Zhijun Gao[3,*], Shipian Shao[1] and Xinyuan Zhang[1]

[1]School of Management, Shenyang Jianzhu University, Shenyang, 110168, China

[2]School of Information and Communication Engineering, North University of China, Taiyuan, 030051, China

[3]School of Information and Control Engineering, Shenyang Jianzhu University, Shenyang, 110168, China

*Corresponding Author: Zhijun Gao. Email: gzj@sjzu.edu.cn

Received: 01 November 2022 Accepted: 13 February 2023 Published: 28 June 2023

ABSTRACT

Building indoor dangerous behavior recognition is a specific application in the field of abnormal human recognition. A human dangerous behavior recognition method based on LSTM-GCN with attention mechanism (GLA) model was proposed aiming at the problem that the existing human skeleton-based action recognition methods cannot fully extract the temporal and spatial features. The network connects GCN and LSTM network in series, and inputs the skeleton sequence extracted by GCN that contains spatial information into the LSTM layer for time sequence feature extraction, which fully excavates the temporal and spatial features of the skeleton sequence. Finally, an attention layer is designed to enhance the features of key bone points, and Softmax is used to classify and identify dangerous behaviors. The dangerous behavior datasets are derived from NTU-RGB+D and Kinetics data sets. Experimental results show that the proposed method can effectively identify some dangerous behaviors in the building, and its accuracy is higher than those of other similar methods.

KEYWORDS

Human skeleton; building indoor dangerous behaviors recognition; graph convolution network; long short term memory network; attention mechanism

Nomenclature

N	Number of joints
T	Number of frames
E	Edge set
H	Set of joint human points
w	Weight function
d	Distance
L	Relative distance
l_{kj}^{t}	Average distance
i	Input gate and
o	Output gate

f	Forgetting gate
b	Deviation
$^{*}g$	Convolution operation
δ	Feature vector

1 Introduction

A dangerous human behavior system is an important part of building an intelligent security system [1], which is crucial for crime prevention and accident handling. According to statistical analysis, modern people spend more than 80% to 90% of their time indoors in a day [2]. With the increasingly prominent problem of population aging in the world and the change in office mode in the post-epidemic era, home care and home office mode will also be the inevitable choices for most people for a long time in the future [3]. Motion recognition plays a crucial role in video surveillance, smart home, elderly and child care, medical care and other applications. It is a hot research issue in the industry and academia at home and abroad [4,5]. With the improvement of people's safety awareness and the demand to build a safe and harmonious society, more and more monitoring equipment are used in banks, supermarkets and other public places.

In the field of human behavior recognition, numerous studies have shown that the use of human skeleton information in feature extraction can effectively overcome the interference of occlusion, motion perspective transformation, complex background, and other factors in the recognition process [6]. With the continuous improvement of hardware devices, the use of human pose recognition technologies such as OpenPose and AlphaPose can directly obtain motion information [7], including the optical flow pose of human skeleton, which can improve the efficiency of related work. Therefore, the human skeleton has received widespread attention because of its strong robustness to the environment [8].

At present, domestic and foreign scholars have conducted extensive research on skeleton behavior recognition methods. Zhao et al. [9] encoded the spatio-temporal information of the skeletal joint point sequence into the view-invariant skeleton mapping and used the three-dimensional convolutional neural network (CNN) to extract features for three-dimensional action recognition. Liu et al. [10] proposed a long short-term memory (LSTM) model with trust gates, which combined the advantages of a long short-term network for long video extraction and reduced the noise of joint point data. Zhou et al. [11] proposed a long short-term network behavior recognition framework based on spatio-temporal convolution, which integrated deep spatial information into each segment to improve recognition performance. Su et al. [12] proposed a multi-loop network fusion behavior detection model to increase the temporal feature extraction ability of the skeleton sequence. Hu et al. [13] proposed a dual attention-guided multi-scale dynamic aggregate graph CNN to establish effective associations amongst human bone nodes to obtain detailed information. Tang et al. [14] applied the deep progressive reinforcement learning method during the extraction of video key frames based on the graph CNN. You et al. [15] extracted the spatial position of key points from the original image of behavior and constructed the skeleton features of human behavior. Then, the edge algorithm is used to extract the skeleton features of different time scales on each edge node and identify them. Although the existing skeleton behavior detection methods have been widely used, the comprehensive extraction of the spatiotemporal features of video data is still difficult. Furthermore, human behavior recognition belongs to inter-class recognition, and the use of salient features is insufficient. In addition to the differences amongst different types of actions, the same kind of action will also have certain differences due to various action ranges considering that the human behavior recognition process

belongs to inter-class recognition [16]. Although the CNN model is relatively simple, problems, such as gradient explosion or gradient disappearance, exist in the actual training process amongst several common neural networks. Although the LSTM network has advantages in extracting video sequences, the extraction effect of spatial features is general, whilst the graph convolution network has a strong ability to capture spatial nodes.

Aiming at the aforementioned problems, this paper proposes a LSTM-graph convolutional network (LSTM-GCN) with an attention mechanism (GLA) model to identify the dangerous behavior of the human body in indoor video surveillance. The model connects GCN in series to the LSTM network so that the skeleton sequence that contains spatial information extracted by the GCN layer can be directly inputted to the LSTM layer for temporal feature extraction and then to realize the direct spatiotemporal fusion. The main innovations of this paper are presented as follows:

(1) A LSTM-GCN with an attention mechanism dangerous behavior recognition model based on attention mechanism that can effectively classify and recognize dangerous behaviors is proposed.

(2) The attention mechanism is integrated into the basic framework of the LSTM-GCN network, which enables the network model to make full use of the salient features and key frames in the video data and solves the problem that the key feature extraction is discontinuous and incomplete in the process of skeleton feature extraction.

(3) The LSTM network is improved in structure, and the bidirectional-LSTM (Bi-LSTM) is used to extract the temporal features of the association between a complete action fully, thereby overcoming the problem of poor long-term memory of the graph convolution network and improves the recognition accuracy. Therefore, this paper introduces Bi-LSTM network to judge the dangerous behavior of the human body.

2 GCN-LSTM Network with Attention Mechanism

The OpenPose open-source programme is a human posture recognition project developed by the Carnegie Mellon University in the United States [17]. It is an open-source library based on CNN and supervised learning and developed with caffe as the framework. It is characterised by capturing the whole-body joints of the human body only through two-dimensional images. Therefore, this study uses OpenPose open-source programme to extract the pixel coordinates of human joint points in the image and matches the pixel coordinates of joint points projected in multiple cameras.

The dangerous behavior recognition detection model proposed in this paper is shown in Fig. 1. The graph convolution layer is used to highlight the extraction of spatial features. The OpenPose toolbox is used for target locking and bone point extraction. Its focus is on processing the position change information of bone points in continuous video frames, and the video frames with bone point information are inputted into the GLA network for spatiotemporal feature extraction. The LSTM layer is used to highlight the extraction of time information features, with a focus given on processing the information of bone point data that changes with time points in continuous video frames. Next, the skeletal sequence with spatio-temporal features is inputted into the attention layer to enhance the key features. Finally, Softmax is used to identify and classify building indoor dangerous human behavior.

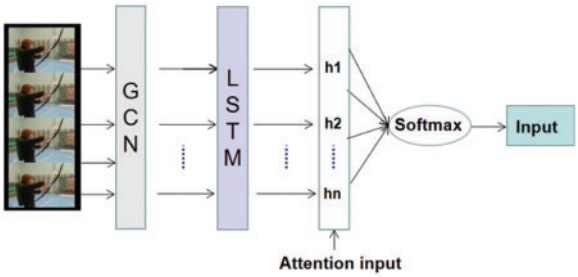


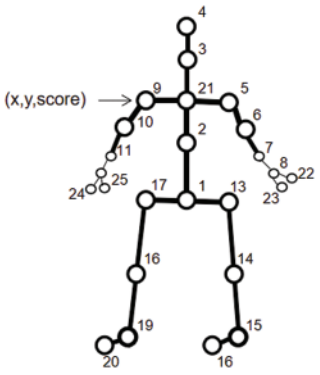
Figure 1: GLA network model structure diagram

2.1 GCN Network Layer

GCN is good at processing some graph data connected by nodes and between nodes, and behavior recognition based on the human skeleton is exactly the same [18]. Firstly, the continuous video is cut into continuous pictures, and then the feature extraction is performed by graph convolution after obtaining the bone information. The human's overall motion information can be represented by joint points. The skeleton is connected by adjacent joint points. Each joint point has a unique coordinate. Subsequently, the motion state of the human body is analyzed by the information of these coordinate changes. Thus, the network input corresponds to the joint coordinate vector of the graph node, similar to an image-based CNN, and then generates a deeper feature input to the next network layer.

2.1.1 Model Architecture

The OpenPose toolbox can be used to obtain the coordinate data of human joint points, and the number of joint points extracted is 25 (Fig. 2). The number of each joint point in the graph is unique, and its feature vector consists of three parts, namely, x, y, and confidence score.



Number	Joint	Number	Joint	Number	Joint
1	SpineBase	10	ElbowLeft	19	AnkleLeft
2	SpineMid	11	WristLeft	20	FootLeft
3	Neck	12	HeadLeft	21	SpineShoulder
4	Head	13	HipRight	22	ThumbRight
5	ShouldRight	14	KneeRight	23	HandTipRight
6	ElbowRight	15	AnkleRight	24	HeadTipLeft
7	WristRight	16	FootRight	25	ThumbLeft
8	HeadRight	17	HipLeft		
9	ShouldLeft	18	KneeLeft		

Figure 2: Openpose skeleton collection diagram

The human motion process is represented by dividing the video frame into multiple separate video frame images, and the joint points will change in time and space accordingly. Let the video frame that contains the human skeleton be represented as $G = (V, E)$ the number of joints in the picture is N, the number of frames of the video is T, the set of the entire joint point sequence is $V = \{v_{ti} | t = 1, \ldots, T; i = 1 \ldots N\}$, and the edge set is E, which contains time and space. In a separate video frame, the joints are connected by edges according to the structure of the human body as

$E_s = \{v_{ti}v_{tj}|(i,j) \in H\}$, and H is the set of joint human points. The connection amongst different frames of the same joint is expressed as $E_F = \{v_{ti}v_{(t+1)j}|(i,j) \in H\}$.

2.1.2 GCN Network Layer Structure

After obtaining the human skeleton data, a pixel matrix with the same size as the weight matrix near each node position is extracted. The feature vectors on these pixels are spliced in spatial order and inner-producted with the parameter vector of the convolution kernel to obtain the convolution output value at this position. The schematic diagram is shown in Fig. 3:

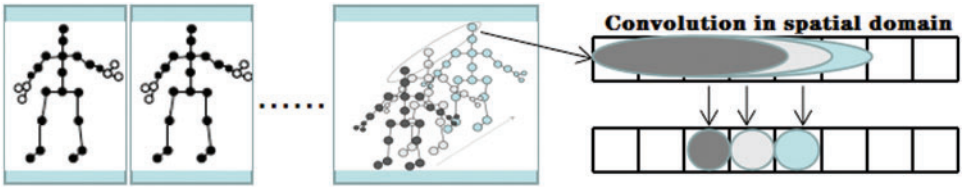


Figure 3: Convolution graph on GCN spatial domain

The step size is set to 1, the convolution kernel size is set to $K * K$, f_{in} is the mapping of the input feature, and then the spatial position of the single channel at x is outputted as shown in Eq. (1):

$$f_{out}(x) = \sum_{h=1}^{K}\sum_{w=1}^{K} f_{in}(p(x,h,w)) * w(h,w) \tag{1}$$

The sampling function is that $p(x,h,w)$ is a pixel point of $K * K$ size around the center position, which is used to determine the position of the subset in the neighborhood. The neighborhood value is set to 1, and the weight function is w. Its function is to assign values to the convolution kernel, and its input value is fixed to achieve the sharing of convolution kernel weights and make the sampled pixels the same. Based on this, Eq. (1) is extended to the human skeleton map, and the equation is now redefined.

Let $B(v_{ti}) = \{v_{ti} | d(v_{tj}, v_{ti}) \le D\}$ be the neighborhood of node v_{ti}, where $d(v_{tj}, v_{ti})$ is the minimum distance between b and c, and the maximum value of the neighborhood range is set to 1. Hence, the sampling function is:

$$p(v_{ti}, v_{tj}) = v_{ti}, v_{tj} \in B(v_{ti}) \tag{2}$$

In the general two-dimensional convolutional network structure, the sampling unit will select the adjacent pixels according to the spatial structure of the two-dimensional grid, that is, when the grid is fixed in the center position, the adjacent pixels will have a relatively fixed arrangement as the position. In the structural diagram of the human skeleton, the spatial structures such as grids, the number of nodes, and the order of arrangement are different. Therefore, the neighborhood range of the vertex must be set artificially, the neighborhood nodes should be divided into K fixed subsets and must number them as a unified mapping $L_{ti}: B(v_{ti}) \to \{0, \ldots, K-1\}$. Thus, we can assign a weight function to this subset, as expressed as Eq. (3):

$$w(v_{ti}, v_{tj}) = w'(\ln(v_{tj})) \tag{3}$$

In the space division strategy, it is divided into three parts, namely, the root node itself, the centripetal set and the centrifugal set. The centripetal set represents the adjacent nodes closer to the

bone center than the root node, and the centrifugal set represents the adjacent nodes farther away from the bone center than the root node. The expression is presented as follows:

$$L\left(v_i, v_j\right) = \begin{cases} 0, & r_j = r_i \\ 1, & r_j < r_i \\ 2, & r_j > r_i \end{cases} \tag{4}$$

where $L_{(v_i,v_j)}$ represents the relative distance between two nodes; r_i and r_j represent the relative distance between two skeletal points and the center of gravity of the human body, respectively. Finally, a graph convolutional layer network model is constructed to extract features from the data. The equation is expressed as Eq. (5):

$$f_{out}\left(v_i\right) = \sum_{v_j \in B\left(v_i\right)} \frac{1}{Z_{i,j}} f_{in}\left(v_j\right) * w\left(L\left(v_i, v_j\right)\right) \tag{5}$$

where x is the weight, y is the normalisation term to balance different subsets, and z represents the corresponding number of nodes in the subset.

2.1.3 Joint Point Weight Algorithm

All the joints of the human body involved in modelling during the recognition process affect the accuracy of the results, and their contribution also varies. The actual detection found that the hand, head, foot, and other joints in the recognition process have more obvious effect. They all have one thing in common, that is, the distance from the center of gravity of the body is far. This article selects the chest bone point as the center of gravity. The line from the coordinate $\left(x_{ji}, y_{ji}\right)$ of the remaining bone points to the center of gravity is defined as l. Higher scores are given to those with a larger l value. Here, we take the average of the distance sum of the continuous $(t, t + \Delta t)$ moments and deal with its specific calculation method, as shown in Eq. (6):

$$\vec{l}_{kj}^{t} = \frac{\left(\left|l_{kj}^{t}\right| + \left|l_{kj}^{t+\Delta t}\right|\right)}{2} \tag{6}$$

l_{kj}^{t} represents the average distance from the remaining bone point to the center of gravity. The larger the value, the farther the bone point is from the center of gravity, and a small value indicates that the distance from the center point is close, thereby further highlighting the role of the key bone points. In this way, the confidence score of each joint point is different, the representation of the feature vector will also be different and the distinction among different actions will become evident.

2.2 LSTM Network Layer

The LSTM network refers to an improved recurrent neural network that is often used to train dynamic time series models [19]. The LSTM network considers the relationship between the output of the current moment and the output of the previous moment, which is more conducive to extracting hidden information and is not prone to problems such as gradient disappearance and explosion.

In fact, the GCN relies on the expansion and multi-layer stacking of convolution kernels to improve the receptive field in the time domain in the feature learning process, which restricts the ability to learn long-term features. Therefore, this paper uses Bi-LSTM.

2.2.1 Input Mode of the LSTM Layer

The general LSTM unit contains input gate, forgetting gate and output gate. Now, we take the output of the GCN layer as input to the LSTM (Fig. 4).

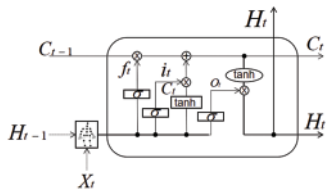


Figure 4: LSTM unit under convolution input

The LSTM recursion under the graph convolution input is shown in Eqs. (7)–(11):

$$f_t = \sigma\left(W_{xf*g}X_t + W_{\ln f*g}H_{t-1} + bf\right) \tag{7}$$

$$i_t = \sigma\left(W_{xi*g}X_t + W_{hi*g}H_{t-1} + b_i\right) \tag{8}$$

$$o_t = \sigma\left(W_{xo*g}X_t + W_{ho*g}H_{t-1} + b_o\right) \tag{9}$$

$$C_t = f_t c_{t-1} + i_t \tanh\left(W_{xc*g}X_t + W_{hc*g}H_{t-1} + b_c\right) \tag{10}$$

$$Ht = ot \odot \tanh(ct) \tag{11}$$

The model has two activation units, namely, the input and output activation, where tanh function is the activation function; f, i and o are the input gate, output gate and forgetting gate, respectively. Matrix W is the weight, variable b is the deviation, and $^*g$ is the convolution operation. In this way, the hidden layer and memory unit connected by GCN and LSTM contain two-layer features in the spatiotemporal domain.

2.2.2 Bi-LSTM Layer

The Bi-LSTM network has two network layers (the structure is shown in Fig. 5), which learn deeper temporal features from forward and reverse. The structural improvement also enhances the spatial features of the graph convolutional layer input.

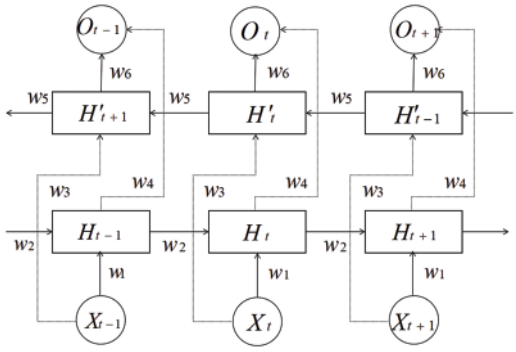


Figure 5: Bi-LSTM structure diagram

It operates in the same way as before, but consists of two parts when processing timing characteristics, forward postprocessing and backward forward processing. The equation is expressed as Eqs. (12)–(14):

$$H_t = f\left(w_1 X_t + w_2 H_{t-1} + b_c\right) \tag{12}$$

$$H'_t = f\left(w_3 X_t + w_5 H_{t-1} + b_e\right) \tag{13}$$

$$O_t = w_4 H_t + w_6 H'_t \tag{14}$$

The bidirectional structure combines the time dynamics of the circulatory system for feed forward and backward training models. In the training process of the Bi-LSTM structure, the Bi-LSTM calculates two sequences (e.g., forward hidden sequence and backward hidden sequence) to generate output sequence by iterating forward layer rising from time t = 0 to t = T and hidden backward layer falling from time t = T to t = 1. The network layer specification is shown in Fig. 6.

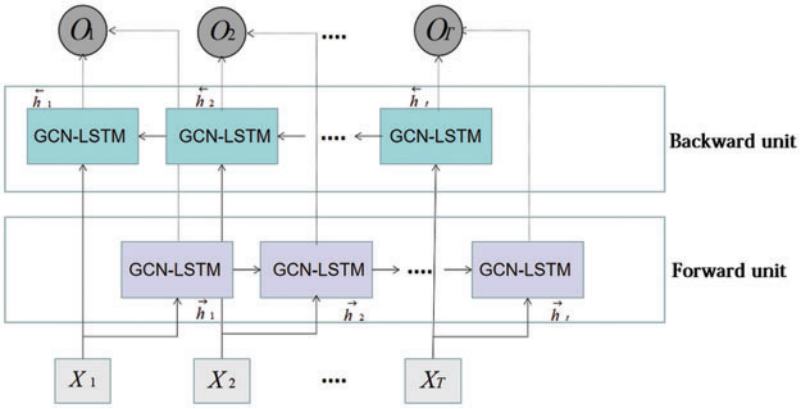


Figure 6: Single-layer Bi-LSTM network structure

3 Attention Layer

The attention mechanism suggests that people or machines selectively focus on and process information with different degrees of importance according to their needs [20]. It has been widely used in the field of computer vision. It automatically analyses the important part of the data characteristic data to predict the relationship between the results in order to facilitate some of the default input and output debugging parameters. The purpose of adding attention mechanism is to enlarge some key features. For the skeletal behavior recognition in this paper, the key is the extraction of skeletal features. This paper adds an attention layer after the GCN-LSTM layer (Fig. 7).

After entering the attention layer, the attention scores of different sequences are added, and the equation is expressed as Eq. (15):

$$\hat{O}_t = f_{att}\left(O_t\right) + O_t \tag{15}$$

This strengthens the key skeleton points without weakening the information for all points, as shown in Fig. 8.

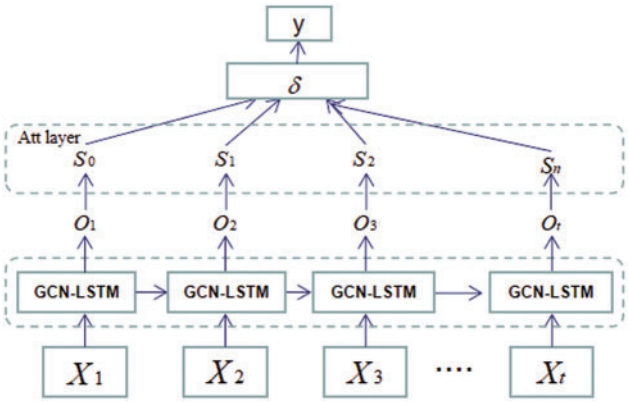


Figure 7: Model diagram incorporating attention mechanism

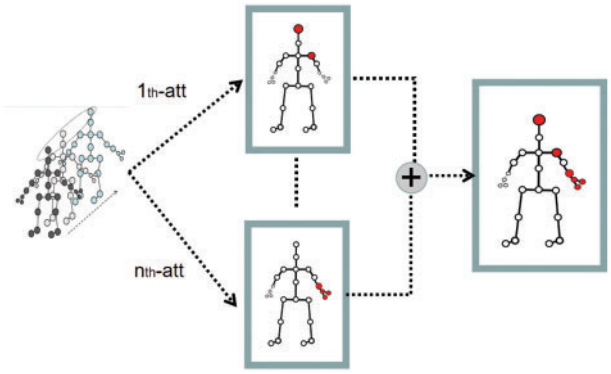


Figure 8: Attention overlay diagram

Next, focus on $\hat{O}_t$ as the output of the GCN-LSTM, as follows:

$$e_t^{(n)} = \tanh\left(We_i \hat{O}_t + b_t\right) \tag{16}$$

$$S_i = \frac{\exp\left(e_t^n\right)}{\sum_{i=1}^{t} \exp\left(e_{i,t}^n\right)} \tag{17}$$

$$\delta = \sum_{i=1}^{n} S_i \hat{O}_i \tag{18}$$

Eq. (17) shows the proportion of the current hidden layer output to the overall score. The greater the value, the greater the overall 'attention' is in the whole at the current time to realise the conversion of the current original hidden layer state to the attention state, and the feature vector δ that contains the skeleton sequence is calculated by Eq. (18). Finally, Softmax is used to integrate and complete the classification and identification of dangerous behaviors.

4 Experimental Results and Analysis

4.1 Dataset

The dataset of indoor dangerous behaviors used in this paper is selected from NTU-RGB+D dataset. NTU-RGB+D dataset is a widely used public 3D human behavior dataset at present. This dataset is collected by three Kinect2.0 cameras, which the same height but have different horizontal angles, namely, −45°, 0° and 45°. Therefore, the author of this dataset presents two reference evaluation benchmarks, namely, cross subject (CS) and cross view (CV). The CS is collected by more than one volunteer for the same action in the training and test sets. CV tests the robustness of the method from three different perspectives. This paper uses the samples collected by camera 1 as the test set and those by cameras 2 and 3 as the test set.

Kinetics dataset is a widely used dataset in the field of human behavior recognition. It is composed of all the key points of the human skeleton in the video identified by Google DeepMind through OpenPose pose estimation software. The dataset divides actions into 400 categories, with a total of about 300,000 videos. To match the number of people in the NTU-RGB+D dataset, the two highest skeletons are selected for action videos with more than two people. To facilitate the calculation of some low confidence video clips deleted, they are not included in the final result. The accuracy of the [21] Kinectics dataset is compared with the accuracy of the Top 1 and Top 5.

Considering that the Kinetics dataset contains many types of actions, this paper selects 15 actions that intersect with the NTU-RGB+D dataset as dangerous actions. These 15 dangerous actions are divided into single-person actions and double-person actions. Between the two classifications dangerous single-person actions are defined as kicking things, back pain, neck pain, vomiting, falling, headache, nausea, throwing things, chest pain, and hobbling. Double actions including pushing, pointing at someone, slapping, kicking and touching a wallet, are defined as dangerous. Approximately 16,000 video clips, two-thirds and one-third of which account for the training set and test set, respectively. Another 45 actions are taken to define non-dangerous actions. The training set and test set ratio of approximately 40,000 video clips are consistent with the aforementioned number. The Resize function is used to standardise images to 224 × 224 when processing video frames.

4.2 Experiment and Result Analysis

To verify the performance of the GLA network in dangerous behavior recognition, we have performed three groups of experiments, including accuracy verification experiment, attention mechanism visualisation experiment and GLA network and other algorithm comparison experiments.

4.2.1 Precision Verification Experiment

The language used in this method is Python, and the deep learning framework is TensorFlow2.0. In the training optimisation algorithm, this paper selects the Adam optimisation algorithm. In this paper, the parameters of the Bi-LSTM model are selected from the empirical data, and Dropout is set to 0.5 to prevent overfitting. The initial learning rate was set to 0.01, the batch-size of NTU-RGB+D was set to 64, and the batch-size of the Kinetics dataset was set to 32. In the GLA network, the dimension of the hidden state and memory unit is 256, and the maximum number of network layers is 3. The NTU-RGB+D dataset training result graph is shown in Fig. 9. Fig. 9a illustrates the training accuracy curve of the dataset. The test accuracy of the model on the NTU-RGB+D dataset is 81.4% (CS) and 89.6% (CV). Fig. 9b presents the loss rate curve. The graph shows that it has good adaptability and faster convergence speed for the NTU-RGB+D dataset. The reason is that after integrating the

attention mechanism, the model can learn more important feature information and has a stronger ability to distinguish different actions with minimal changes in joint point coordinates.

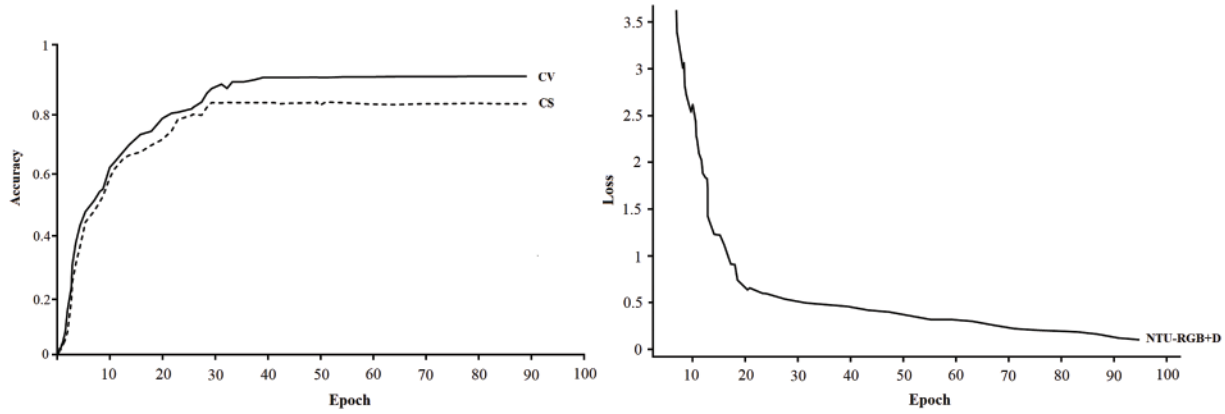


Figure 9: NTU-RGB+D dataset training result graph: (a) accuracy; (b) loss

In addition, the number of GLA network layers will also affect the accuracy of identification. Thus, this paper conducted a number of comparative experiments. In the original GCN-LSTM network, the network layer is set to 2, 3, 4, 5 and 6 layers. However, in the GCN-Bi LSTM network, the two-way long and short term has been divided into two layers. Here, the network layer, which is divided into 2 and 3 layers, is merged and tested. Now, these two different ways of connection are compared. Only the comparison of the accuracies of CS and CV is performed here. The accuracy is shown in Table 1:

Table 1: Accuracy comparison of different network layers

Model	Layer number	CS%	CV%
GCN-LSTM	2	81.2	87.7
	3	82.9	89.6
	4	85.6	91.2
	5	88.3	91.0
	6	84.7	90.9
GCN-Bi LSTM	2	85.3	92.2
	3	89.1	93.7

The experimental results show that as the number of network layers in the GCN-LSTM network increases, the accuracy also improves first and then decreases. The reason is that the ability of the model to extract features will be enhanced with the increase of the number of network layers. However, the saturation of gradient differences and the difficulty in training and optimization will also increase. Compared with the single-layer LSTM network and the double-layer multi-directional LSTM network, the accuracy of the Bi-LSTM model is higher than that of the original model under the same number of network layers, thereby also showing that the improved Bi-LSTM network layer has stronger ability to extract skeleton features.

In the Kinetics dataset verification experiment, referring to the comparison method in the NTU-RGB+D dataset, the highest CV accuracy is selected from the three deep learning-based methods for repeated verification. The accuracy indicators are Top 1 and Top 5, and the comparison results are shown in Table 2.

Table 2: Accuracy comparison of different network layers

Algorithm	Top1	Top5
AGCN	37.8	59.3
Res-CNN	37.3	58.9
Ind-RNN	38.1	59.8
LSTM-GCN	38.4	60.3

The simulation results show that the model in this paper still performs best in the test of the Kinetics dataset, which further proves the universality of the method in this paper.

4.2.2 Visualisation of Attention Mechanism

The results were compared visually to verify the necessity of incorporating attention mechanism into this study. Fig. 10 represents the attention distribution vector of the kicking action, where (a), (b) and (c) correspond to different input attention vectors, and the size of the blue circle represents the distribution of the attention values. The larger the circle, the greater the corresponding attention value will be. Fig. 10a represents the concerned right hip and right knee. Fig. 10b represents the right ankle and right foot segments of concern. Fig. 10c represents both of them. Interestingly, the distribution of attention fits the characteristics of kicking. Fig. 11 shows the attention distribution vector of the action. Without the loss of generality, it also has better joint point recognition ability in double multi-body movements, as shown in Fig. 11.

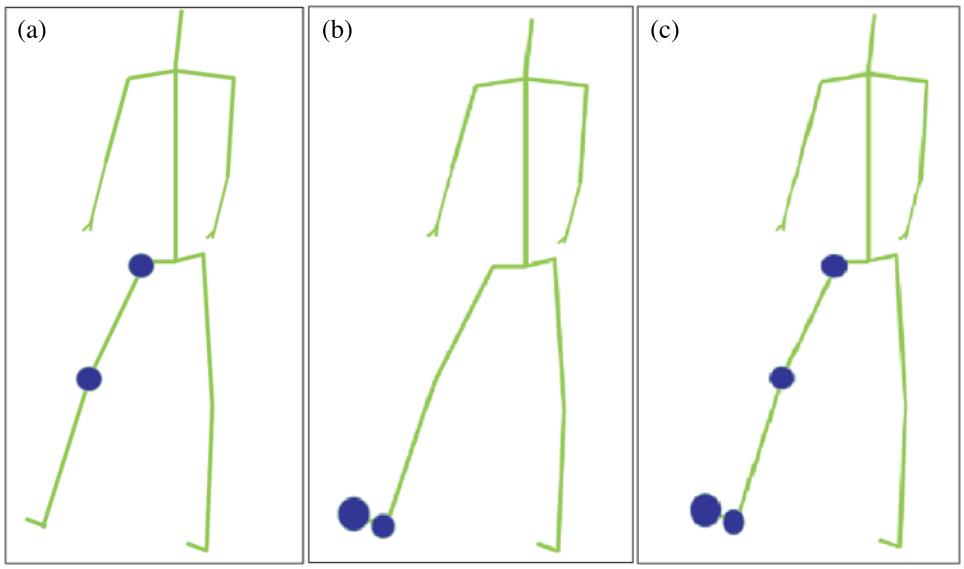


Figure 10: Attention distribution vector of action 'kick'

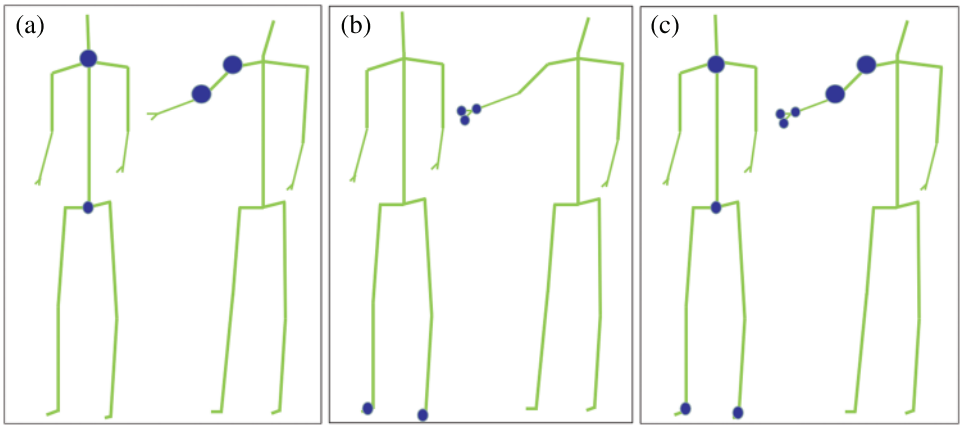


Figure 11: Attention distribution vector of multi-action 'finger'

4.2.3 Comparative Computational Experiments

As a commonly used visualization method to present the performance of the algorithm in the field of scene classification, the confusion matrix can measure the confusion degree of the classification model intuitively and concisely [22]. In the confusion matrix, the values on the diagonal are all correct prediction results, and the remaining values are the wrong prediction results caused by model misjudgement [23]. Each row of the matrix represents the real category, and each column of the matrix represents the prediction label of the model. The confusion matrix obtained in this paper is shown in Fig. 12, and the classification accuracy of the drop is 100%.

Predictive behavior categories \ Real behavior category	kicking things	back pain	neck pain	vomiting	falling	headache	nausea	throwing things	chest pain	hobbling	pushing	pointing at someone	slapping	kicking	touching a wallet
kicking things	0.958	0	0	0	0	0	0	0	0	0.132	0	0	0	0.116	0
back pain	0	0.961	0.063	0	0	0	0	0	0	0	0	0	0	0	0
neck pain	0	0.087	0.884	0.094	0	0.075	0	0	0	0	0	0	0	0	0
vomiting	0	0	0.128	0.824	0.027	0	0	0	0	0	0.114	0	0	0	0
falling	0.051	0	0	0.024	1	0	0	0	0	0.121	0.052	0	0	0	0
headache	0	0	0.073	0	0	0.853	0	0	0	0	0	0	0	0	0
nausea	0	0	0	0	0	0	0.796	0	0	0	0	0	0	0	0
throwing things	0	0	0	0	0	0	0	0.914	0	0	0.069	0.084	0.059	0	0.031
chest pain	0	0.091	0	0.036	0	0	0.038	0	0.885	0	0	0	0	0	0
hobbling	0.102	0.116	0	0.072	0.119	0	0	0	0.031	0.710	0	0	0	0	0
pushing	0	0	0	0	0	0	0	0.014	0	0	0.809	0.045	0.142	0	0.087
pointing at someone	0	0	0	0	0	0	0	0.005	0	0	0.117	0.682	0.139	0	0.740
slapping	0	0	0	0	0	0	0	0.021	0	0	0.109	0.094	0.942	0	0.309
kicking	0.132	0	0	0	0	0	0	0	0	0.083	0	0	0	0.768	0
touching a wallet	0	0	0	0	0	0	0	0	0	0	0.075	0.129	0.113	0	0.837

Figure 12: Confusion matrix

In addition, to prove the superiority of the GLA network, its accuracy is compared with that of some mainstream methods.

The results in the Table 3 show that the proposed GLA model in this paper has certain advantages and reference significance over other popular methods.

Table 3: Accuracy comparison of different algorithms

Algorithm	CS%	CV%
3D-CNN [24]	79.6	84.8
Visaualize-CNN [25]	76.0	82.6
Two stream-LSTM [26]	81.8	89.6
ATT-ConLSTM [27]	76.1	84.0
ST-AGCN [28]	86.4	92.1
HCN [29]	86.5	91.1
GLA	89.1	93.7

5 Conclusion

A LSTM-GCN network model that incorporates the attention mechanism is proposed to identify dangerous behaviors based on the indoor conditions of buildings. The model connects GCN and LSTM networks in series. Firstly, the bone information is inputted into the GCN network for spatial feature extraction. Secondly, the bone sequence that contains spatial information is inputted into the LSTM layer for temporal feature extraction to achieve direct spatio-temporal fusion. Bidirectional long short-term network is used to extract temporal features from two directions, which overcomes the problem of poor long-term memory of GCN in feature learning. Thirdly, we designed an attention layer to enhance the key skeletal information. Finally, the simulation results show that the proposed method can effectively achieve the expected goal and has certain practical application value. At the same time, considering the dangerous behavior recognition is 'afterwards' recognition research, that is, the system needs to identify the action after it is completed. In many application scenarios, people prefer to predict and identify the action in time before the completion of the action, which can provide us with enough reaction time to prepare in advance. For example, when the system observes that a person has lost his balance, he may fall. We hope that the system can predict the occurrence of this action in time, take appropriate measures in advance and respond accordingly to avoid irreparable accidents.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Zhu, J. B., Xu, Z. L., Sun, Y. W., Ma, Q. S. (2020). Detection of dangerous behaviors in power stations based on OpenPose multiperson attitude recognition. *Automation & Instrumentation, 35(2),* 47–51.

2. Rupp, R. F., Vásquez, N. G., Lamberts, R. (2015). A review of human thermal comfort in the built environment. *Energy and Buildings, 105,* 178–205.
3. United Nations, Dept of International Economic and Social Affairs (2002). *World population ageing: 1950–2050.* New York: United Nations.
4. Wang, L., Koniusz, P., Huynh, D. (2019). Hallucinating IDT descriptors and I3D optical flow features for action recognition with CNNs. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8698–8708. Seoul, Korea (South).
5. Yan, S., Xiong, Y., Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. *AAAI Conference on Arfificial Intelligence*, pp. 7444–7452. New Orleans.
6. El Kaid, A., Baïna, K., Baïna, J. (2019). Reduce false positive alerts for elderly person fall video-detection algorithm by convolutional neural network model. *Procedia Computer Science, 148(2),* 2–11.
7. Guo, Y., Guo, C. X., Shi, X., Shen, L. M. (2020). Adaptability of study desks and chairs based on analysis of sitting posture using OpenPose. *Journal of Forestry Engineering, 5(2),* 179–185.
8. Simonyan, K., Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. *Advances in Neural Information Processing Systems, 27,* 568–576.
9. Zhao, Y., Wen, L., Li, S., Cheng, H., Zhang, C. (2019). A view-invariant skeleton map with 3DCNN for action recognition. *2019 Chinese Automation Congress (CAC)*, pp. 2128–2132. Hangzhou, China.
10. Liu, J., Shahroudy, A., Xu, D., Wang, G. (2016). Spatio-temporal LSTM with trust gates for 3D human action recognition. *European Conference on Computer Vision*, pp. 816–833. Amsterdam, The Netherlands.
11. Zhou, K., Wu, T., Wang, C., Wang, J., Li, C. (2020). Skeleton based abnormal behavior recognition using spatio-temporal convolutional attention based LSTM. *Procedia Computer Science, 174,* 424–432.
12. Su, T. T., Sun, H. Z., Ma, C. M. (2018). Research on human behavior recognition based on recurrent neural networks. *Journal of Tianjin Normal University (Natural Science Edition), 38(6),* 58–62+76.
13. Hu, Z., Lee, E. J. (2020). Dual attention-guided multiscale dynamic aggregate graph convolutional networks for skeleton-based human action recognition. *Symmetry, 12(10),* 1589–1604.
14. Tang, Y., Tian, Y., Lu, J., Li, P., Zhou, J. (2018). Deep progressive rein-forcement learning for skeleton-based action recognition. *Proceedings of the IEEE Conference on European Conference on Computer Vision*, pp. 5323–5332. Salt Lake City, UT, USA.
15. You, W., Wang, X. (2020). Study on the edge computing method for skeleton-based human action feature recognition. *Chinese Journal of Scientific Instrument, 41(10),* 156–164.
16. Nakano, N., Sakura, T., Ueda, K., Omura, L., Kimura, A. et al. (2020). Evaluation of 3D markerless motion capture accuracy using OpenPose with multiple video cameras. *Frontiers in Sports and Active Living, 2,* 50.
17. Liu, Y., Zhang, C., Cheng, J., Chen, X., Wang, Z. J. (2019). A multi-scale data fusion framework for bone age assessment with convolutional neural networks. *Computers in Biology and Medicine, 108(1),* 161–173.
18. Thomas, N. K., Max, W. (2016). Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representation*, pp. 101–111. arXiv:1609.02907.
19. Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9(8),* 1735–1780.
20. Bahdanau, D., Cho, K., Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *International Conference on Learning Representations.* arXiv:1409.0473.
21. Zhang, P., Lan, C., Xing, J., Zeng, W., Xue, J. et al. (2019). View adaptive neural networks for high performance skeleton-based human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(8),* 1963–1978.
22. Bai, Z. H., Wang, Y. Y., Zhang, L. W. (2020). Driver distraction behavior detection with multi-information fusion based on graph convolution networks. *Automotive Engineering, 42(8),* 1027–1033.
23. Wakefield, D. J. (1988). Application of the human cognitive reliability model and confusion matrix approach in a probabilistic risk assesmeent. *Reliability Engineering & System Safety, 22,* 295–312.

24. Ke, Q., Bennamoun, M., An, S., Sohel, F., Boussaid, F. (2017). A new representation of skeleton sequences for 3D action recognition. *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3288–3297. Honolulu, USA.
25. Liu, M., Liu, H., Chen, C. (2017). Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition, 68,* 346–362.
26. Zheng, W., Li, L., Zhang, Z., Huang, Y., Wang, L. (2019). Relational network for skeleton-based action recognition. *Proceedings of 2019 IEEE International Conference on Multimedia and Expo*, pp. 826–831. Shanghai, China.
27. Liu, J., Wang, G., Duan, L. Y., Abdiyeva, K., Kot, A. C. (2018). Skeleton-based human action recognition with global context-aware attention LSTM networks. *IEEE Transactions on Image Processing, 27(99),* 1586–1599.
28. Yi, C., Wu, W., Li, P., Xia, Y., Gao, Q. (2020). Skeleton-based action recognition based on spatio-temporal adaptive graph convolutional neural-network. *Journal of Huazhong University of Science and Technology (Natural Science Edition), 48(11),* 5–10.
29. Li, C., Zhong, Q., Xie, D. (2018). Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 786–792. Stockholm, Sweden.