ISSN 2616-5775 Vol. 5, Issue 6: 89-94, DOI: 10.25236/AJCIS.2022.050613

Real time dangerous action warning system based on graph convolution neural network

Cong Wang, He Zhang, Zhengli Zhai

School of Information and Control Engineering, Qingdao University of Technology, Qingdao 266520, Shandong, China

Abstract: *In order to solve a series of disadvantages of slow efficiency and high cost brought by the traditional human action recognition method based on manual feature extraction, in order to prove the feasibility of using graph convolution neural network and action recognition, a real-time dangerous action alarm system based on graph convolution neural network is designed and implemented. The system includes data collection, human posture estimation, action recognition, dangerous action alarm and other functions. The feasibility of using graph convolution neural network to realize real-time dangerous action alarm system is proved by the system. Even if the graph convolution neural network does not pass the training, it can extract good features only by using the original random initialization parameters. If certain annotation information is given, its effect can be greatly improved. After the system adopts the graph convolution neural network, the action recognition can be more accurate, more efficient, faster and more reliable than the traditional method.*

Keywords: *graph convolution neural network; Action recognition; Dangerous action alarm*

1. Introduction

With the development of visual sensors such as Microsoft Kinect and the improvement of people's security awareness, the application scope of real-time dangerous action alarm system is expanding. Compared with ordinary monitoring, the main difference of real-time dangerous action alarm system is that it embeds human action recognition technology into the monitoring server, uses algorithms to identify human actions in the monitoring screen, and sends an alarm to users in time when abnormal behaviors occur. Rapid and accurate warning of dangerous actions can effectively improve people's perception of abnormal behavior.

Thanks to the development of visual sensors, the improvement of computer computing power and the rapid development of deep learning algorithms, the construction of smart city is entering a more intelligent stage. People put forward higher requirements for improving the effectiveness of urban management and ensuring the quality of life of citizens. Real time risk alarm has broad application prospects in many social and public security fields such as streets, banks, supermarkets, campuses and communities. Take campus scenes as an example, The real-time dangerous action alarm system can use human action recognition to evaluate the psychological status of students, so as to timely prevent the occurrence of campus bullying and ensure that students have a more benign and healthy living environment for learning. With the increase of monitoring equipment, the amount of data to be processed is growing rapidly. It is easy to miss key information only by processing such a large amount of monitoring data by manpower, which not only consumes manpower, but also has low efficiency.

According to the above background, this project plans to embed the graph convolution neural network model into the monitoring system, use the attention mechanism to solve the problems of many joint points and large degrees of freedom in human action recognition, combine the long short term memory (LSTM) unit to integrate the spatial and temporal characteristics of human bones, and learn the multi angle, short-term and long-term characteristic information of action sequence at the same time, To effectively carry out human posture estimation and action recognition, so as to alarm dangerous actions more accurately. For different application scenarios, different definitions of dangerous actions can be made, and the monitoring picture can be captured. After the image is collected, the system transmits the image to the skeleton joint point extractor through the action detector to help the algorithm identify the human actions in the monitoring picture and alarm the dangerous actions, so as to guide the user to carry out further processing in time, alarm the dangerous actions, and guide the user to carry out the next processing in time.

ISSN 2616-5775 Vol. 5, Issue 6: 89-94, DOI: 10.25236/AJCIS.2022.050613

2. Basic knowledge of relevant theories

2.1 Acquisition of skeleton joint points

In the scheme of taking RGB video as the research sample, most researchers only pay attention to the pixel information in the image, ignoring the performance of human limb movements in the process of movement, which is mainly completed by the mutual traction and cooperation between skeleton and joint points. Therefore, human skeleton joint graph contains rich action feature information. However, most motion recognition data sets, such as 20bn jester and dynamics, only contain RGB video or image samples, and there is no human joint point information. Recently, some studies have found that skeleton data can be considered as a relatively high-level feature in human motion, which has strong robustness to scale change and illumination change, and is invariant to the perspective of acquisition equipment, human motion rotation and action speed. Because skeleton data can not be affected by visual self occlusion, background interference and illumination change, joint images with motion space-time features are more robust [1]. The movement of human body can be represented by the changes of skeleton joint data on a series of frames. Each frame has one or more sets of joint point coordinates and connecting edges between joint points. The skeleton joint data is collected from the human motion changes in the scene through the 3D somatosensory motion capture device. The coordinates of the human skeleton joint points in the two-dimensional space can be expressed as (x, y, score), where score is the confidence of the joint points. [1] Ntu-rgbd is the representative data set of commonly used skeleton joint point data, which represents the whole person with 25 main joint points, as shown in Figure 1. Human skeleton joint points form a natural skeleton joint diagram through the connection of biological characteristics, but different nodes have different topological structures, that is, the number of adjacent nodes is different. For example, the central node 31 has four adjacent nodes, which are 3, 9, 2 and 5 respectively, while the 22 nodes of the hand have only one neighbor node 23. The joint points with different topologies form a spatiotemporal skeleton graph, in which the vertices in the skeleton graph are represented by joint points, the time-class edges in the skeleton graph are represented by the sequential connection of the corresponding joint points, and the space-class edges in the skeleton graph are represented by the bone connection of the joint points of the human skeleton.

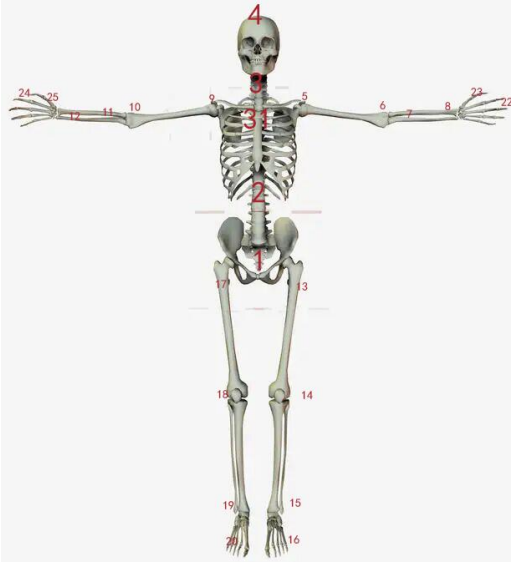

Figure 1: Topological structure of human skeleton in ntu-rgbd

At present, there are mainly two methods to obtain the characteristic information of human joint points in the timing stream. 1) capture the depth information of the movement of the person in the three-dimensional space through Kinect (3D somatosensory camera), and then obtain the coordinates of bone points from the depth map to form the human skeleton joints. 2) the action RGB video stream can use the 2D pose estimation algorithm (such as openpose) to extract the two-dimensional coordinates of the person's joint points and the bone information between joints. This section mainly introduces the first method to obtain skeleton joint points. The data set in the depth image is different from the RGB image to represent the information of intensity or color. Instead, it obtains the depth information from the target scene. Specifically, Kinect is used to locate the position of the target presenter, determine the current foreground and background, and calibrate and collect. Shoton and Fitzgibbon et al. [2] proposed a joint

ISSN 2616-5775 Vol. 5, Issue 6: 89-94, DOI: 10.25236/AJCIS.2022.050613

point extraction method, It can quickly and accurately estimate, predict and obtain the 3D position information of each node from a single depth image.

2.2 Convolutional neural network

In the field of computer, neural network does not refer to the brain neural tissue in the biological concept, but a mathematical model constructed by simulating the information feedback system. Neural network is an adaptive system with simulation and learning functions. According to the data distribution characteristics of different training sets, the internal structure of the model is dynamically changed in the actual training process to make it learn the distribution law of the corresponding data set. At present, convolutional neural network (CNN) is a kind of neural network widely used in various complex tasks, including image and video, including action recognition and video-based target detection [1]. CNN's powerful feature extraction ability and efficient operation speed mainly benefit from its three advantages: local connection, weight sharing and pooling operation. Lecun et al. [3] first proposed to apply CNN model to the field of computer vision in 1998. The model is named letnet5. The model structure is shown in Figure 2. It is used to solve the problem of handwritten digital recognition. Experiments in the paper show that CNN has greatly improved the results over all previous traditional models. Letnet5 model introduces convolution operation, pooling operation and full connection layer, which greatly improves the feature extraction ability of the model.

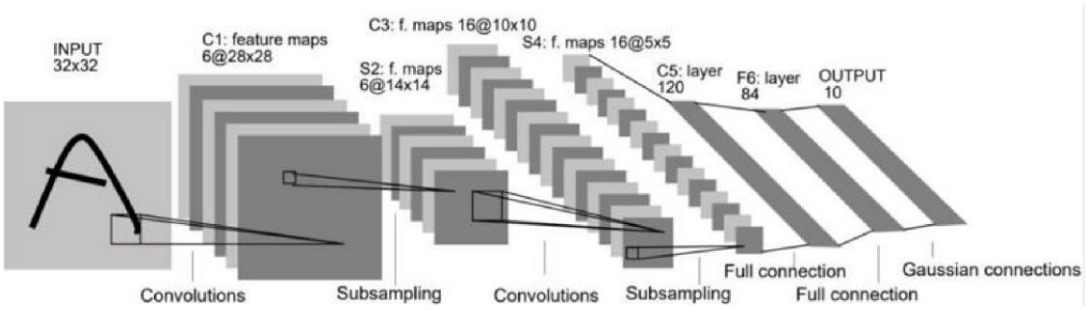

Figure 2: Structure diagram of lenet5 network [3]

2.3 Convolutional neural network

Graph convolution neural network (GCN) is to solve a problem, that is, the current neural network models (such as CNN and RNN) are difficult to extract spatial features from the graph structure for machine learning. The core of convolutional neural network (CNN) lies in its convolution kernel, which can be regarded as a two-dimensional window, and for each picture data, it can be regarded as a two-dimensional grid structure. Therefore, for two-dimensional grid picture data, it is very appropriate to use convolutional neural network to extract features. Recurrent neural network (RNN) is a one-dimensional structure, which is suitable for sequential data such as natural language. Through the operation of various gates, the information before and after the sequence can affect each other (Si, et al. 2019; Liao, et al. 2019; Wang, et al. 2017). Both pictures and natural language belong to European data structure, which is a data structure with limited dimension rules. However, there are many unstructured data in life, such as social network model, point cloud and other data. These data structures are collectively referred to as topology map. In the face of non Euclidean data such as topological graph, which may be infinite dimensions, convolutional neural network and cyclic neural network are difficult to deal with [4].

In order to process the data of topological graph, many new models have emerged, such as GNN, deepwalk, etc., but they all have great limitations. These methods will not only lose a lot of structural information, but also need to rely heavily on graph preprocessing.

Recently, more and more research work has turned to the field of how to design scalable convolutional neural network model to graph data. Most of the research on graph neural network architecture choose to learn from the ideas of other models, including CNN. Shahroudy et al. [5] creatively proposed a modeling method between joints, which draws lessons from the natural connection of human body. Donahue et al. [6] proposed a model gca-lstm combining graph convolution and recurrent neural network. The model can effectively model the global context information in the action sequence, and the model can effectively extract the timing features of human actions. Yan et al. [7] proposed a graph convolution neural network, designed three different sampling functions to construct the graph convolution layer, and combined with the 2D time convolution layer to construct the spatiotemporal graph convolution neural network. The

ISSN 2616-5775 Vol. 5, Issue 6: 89-94, DOI: 10.25236/AJCIS.2022.050613

experiment shows that the model can effectively extract the spatiotemporal features in the joint graph, and achieved good results on the ntu-rgbd motion recognition data set.

Suppose that there are a batch of graph data, in which there are n nodes, and each node has its own characteristics. The characteristics of these nodes are combined into an N×D-dimensional matrix X, each node will form an n N×N The adjacency matrices A, X and A of dimension n are the inputs of the model. The transfer relationship between layers is shown in formula 1:

$$J^{(m+1)} = \sigma\left(\hat{E}^{-0.5}\hat{A}\hat{E}^{-0.5}J^{(m)}W^{(m)}\right) \quad (1)$$

The above formula is the core formula of graph convolution neural network. In formula 1.1, J represents the characteristics of each layer, σ Is an activation function, $\hat{A}$ is the sum of the adjacency matrix A and an identity matrix I of the graph, $\hat{E}$ is the degree matrix of $\hat{A}$, and W is the parameter matrix. Through the observation of formula 1.1, it can be seen that since the numbers on the diagonal of the adjacency matrix A of the graph are all 0, this feature will be ignored when multiplying with matrix J. therefore, it is necessary to add an identity matrix I to the adjacency matrix A. A symmetric and normalized matrix can be obtained by multiplying $\hat{A}$ by the degree matrix. The strength of GCN is that even if the model does not pass training, it can extract good features only by using the original random initialization parameter W. after giving some annotation information, its effect will be greatly improved.

3. System architecture

The architecture design of the action alarm system is shown in Figure 3. The action recognition system is composed of four layers: Web interaction layer, business layer, algorithm layer and data layer.

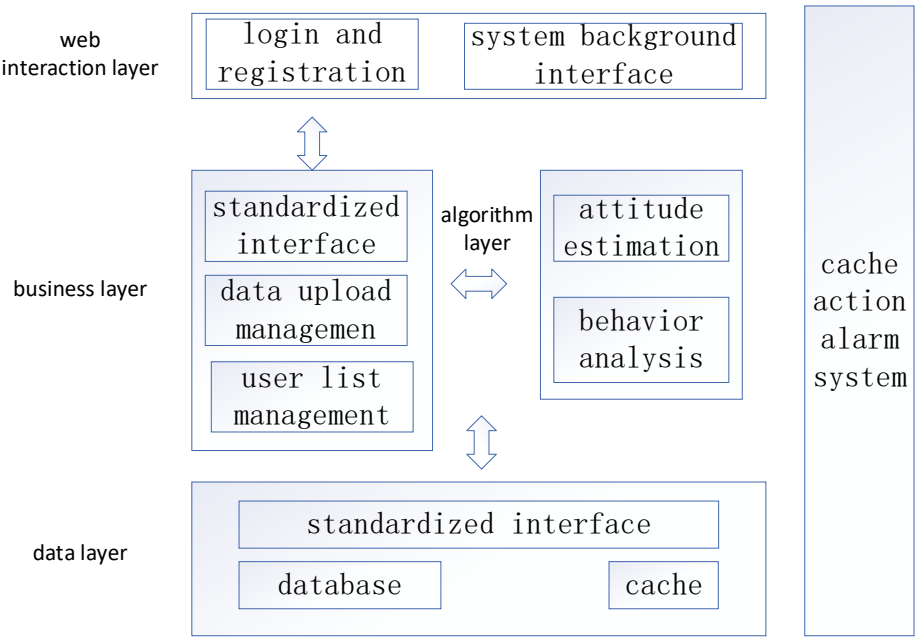

Figure 3: Overall system architecture

The web interaction layer is the interface of the system; The business layer is responsible for handling the relevant business logic existing in the system, mainly providing functions such as user login and registration, video upload and storage, video action recognition analysis and user historical data query. The core part is action recognition analysis. The business layer identifies by calling the algorithm layer, returns the results to the business layer, then transmits them to the web interaction layer, and finally stores them in the data layer; The algorithm layer uses the action recognition algorithm realized by graph convolution neural network to call the data for behavior analysis and write the prediction function into the function, so that the business layer can call the function directly and write the prediction and early warning results into the database; The data layer is responsible for storing user data and caching. The trained neural network model is called to complete attitude estimation, action recognition and dangerous action alarm. The processing of visual data such as pictures and videos is realized, and the output results

ISSN 2616-5775 Vol. 5, Issue 6: 89-94, DOI: 10.25236/AJCIS.2022.050613

are transformed into visual information display. The system is divided into four modules, including data acquisition, attitude estimation, action recognition and dangerous action alarm. The flow chart of real-time dangerous action alarm system is shown in Figure 4.

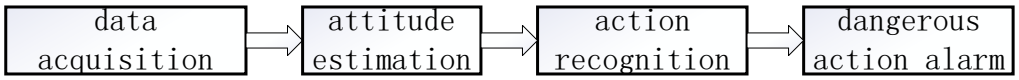

Figure 4: Flow chart of real-time dangerous action alarm system

The design of the main functional modules of the system is shown in Figure 5.

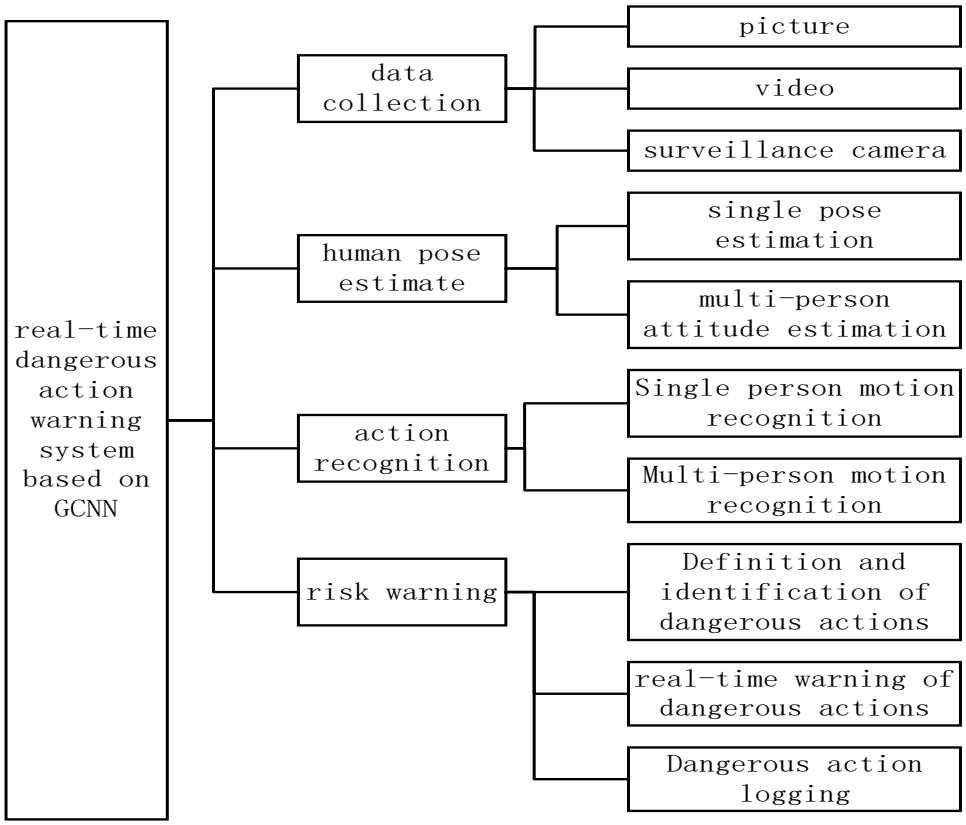

Figure 5: Main function template

4. Conclusion

The graph convolution network with attention mechanism can significantly improve the output effect of the model, the spatial structure information of the data, and complete the end-to-end depth. In addition, the system uses LSTM network to improve the recognition accuracy of associated actions, and can process pictures and videos at the same time. It can be applied to a variety of real scenes such as home, campus, hospital and shopping mall. For example, in families with young children, the system can timely find dangerous actions such as falling and touching electrical appliances; Using this system on campus can effectively prevent the occurrence of campus violence and protect students' physical and mental health. At the same time, the system can define different types of dangerous actions for different scenes through simple customization. Define different types of dangerous actions for different scenes.

References

[1] Ye Dian. Research on human behavior recognition technology based on graph convolution neural network. 2021. Guangdong University of technology, MA thesis.
[2] Shotton J, Fitzgibbon A, Cook M, et al. Real-time human pose recognition in parts from single depth images[C].//CVPR 2011. Ieee, 2011: 1297-1304.
[3] LeC un Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

ISSN 2616-5775 Vol. 5, Issue 6: 89-94, DOI: 10.25236/AJCIS.2022.050613

[4] Yuan Yating. Design and implementation of human motion recognition system based on graph convolution neural network [D]. University of Chinese Academy of Sciences (Shenyang Institute of computing technology, Chinese Academy of Sciences), 2021 DOI:10.27587/d.cnki. gksjs. 2021.000037.
[5] Shahroudy A, Liu J, Ng T T, et al. Ntu rgb+ d: A large scale dataset for 3d human activity analysis[C].//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 1010-1019.
[6] Donahue J, Anne Hendricks L, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 2625-2634.
[7] Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]//Proceedings of the AAAI conference on artificial intelligence. 2018, 32(1).