A Dangerous Behavior Identification System for Public Monitoring

Renkai Hou[1], Yaping Dai[2], and Shuai Shao[*3]

[1]Beijing Institute of Technology, Beijing, China
E-mail: 1030692903@qq.com
[2] Beijing Institute of Technology, Beijing, China
E-mail: daiyaping@bit.edu.cn
[*3] Beijing Institute of Technology, Beijing, China
E-mail: shaoshuai@bit.edu.cn

At the present stage, the identification of dangerous behaviors in public places mostly relies on manual work, which is subjective and has low identification efficiency. This thesis proposes an automatic identification method for dangerous behaviors in public places, which analyzes human behavior and speech emotion through deep learning network and then performed multimodal information fusion. Based on the fusion results, people can judge the emotional atmosphere of the crowd and make early warning and alarm for possible dangerous behaviors. Experiments show that the algorithm adopted in this paper can accurately identify dangerous behaviors and has great application value.

Keywords: Human Behavior Recognition, Speech Emotion Recognition, Multimodal Fusion, Deep Learning

1. INTRODUCTION

With the economic development and social progress, the number of large public places, such as shopping malls and stations, is increasing day by day, and various potential safety problems also follow. If dangerous behaviors cannot be accurately identified when safety risks occur, irreparable losses will be caused. Therefore, people pay more attention to the identification and treatment of dangerous behaviors in public places.

In the process of automatic identification of dangerous behaviors, if we only consider dangerous actions and ignore the emotional factors of human, the results will have great errors. For example, there are dangerous actions when two children are playing, but do not mean dangerous behavior. Therefore, to identify dangerous behaviors more accurately, it is necessary to integrate the information of human action and human emotion. Because voice contains human emotion, the main work of this paper is divided into three parts: human behavior recognition, speech emotion recognition and multimodal information fusion.

Human behavior recognition has been concerned by researchers as early as the 1970s [1]. Early study on human behavior recognition is based on RGB static image or video, and it has low identification accuracy. With the maturity of infrared imaging technology, the depth sensor represented by Microsoft Kinect came out. The device can not only acquire RGB static images, but also acquire depth images and human skeleton data, which brings great convenience to human behavior recognition [2]. Wu used the depth information obtained by Kinect for human behavior recognition. He proposed a moving human body image segmentation method based on an improved two-dimensional Otsu method to solve the problem of inaccurate and slow segmentation of moving human body images using the two-dimensional Otsu method, which increases the accuracy of human body motion tracking by more than 30% [3].

The earliest research on speech emotion recognition appeared in the 1980s, Professor Minsky put forward the idea of "making computers have emotional ability". Aiming at this innovative idea, scholars began to explore speech emotion recognition. Speech emotion feature extraction and feature classification are two key issues in speech emotion recognition. Ghosh extracted the spectrogram in speech to identify the emotion and classified four kinds of emotions through Recurrent Neural Network [4]. Finally, he achieved good results in IEMOCAP.

Multimodal fusion refers to the synthesis of multimodal information, and it can get the final results based on multiple modes by establishing certain algorithms [5]. In the current research on multimodal fusion, the common modes are voice, image and soon. Through the combination of different modes, a variety of problems in different fields can be solved. Sreevidya combined information from audio and video modalities for emotion recognition, and used backpropagation for learning. Both the models are combined by fusion of intermediate layers, and the layers are selected through a grid-based search algorithm [6].

Multimodal fusion method has been successfully applied in many fields, such as image recognition, visual question answering and emotion recognition, and has achieved good results. However, in the field of dangerous behavior identification, the application of multimodal fusion method is comparative little, and the research on this technology is almost blank. Therefore, it is an innovative and breakthrough work to identify dangerous behaviors through multimodal fusion.

2. SINGLE-MODE RECOGNITION FOR HUMAN BEHAVIOR AND SPEECH EMOTION

This paper establishes a dangerous behavior recognition system based on deep learning. Before multimodal fusion, human behavior and speech emotion should be classified respectively, so that their classification results can be obtained. Firstly, the Kinect is used to obtain the coordinates of skeleton points, and the coordinate data is input into the LSTM model to obtain the behavior recognition results. Then, the speech signal is collected through the microphone, and the MFCC feature of the speech is input into the CNN model to obtain the speech emotion recognition results.

2.1. Human Behavior Recognition based on LSTM

The human action in the video is continuous, and the behavior at this moment is not only related to the current position of skeleton points, but also the previous state. Therefore, it is necessary to build a behavior recognition model that can learn in a long-time sequence. During the training process, the traditional neural network model can not process multiple groups of data on the time series at the same time, so it can not solve the problems related to the time series. However, RNN has the ability to solve such problems.

LSTM (Long Short-Term Memory) is essentially a variant or improved form of RNN. Due to the defects of RNN's own structure, there are some problems such as gradient disappearance and gradient explosion. LSTM adds gate structure on the basis of RNN, which can not only deal with time series problems, but also avoid the above problems. Therefore, this paper will use the LSTM network model for human behavior recognition. The network structure is shown in Figure 1.

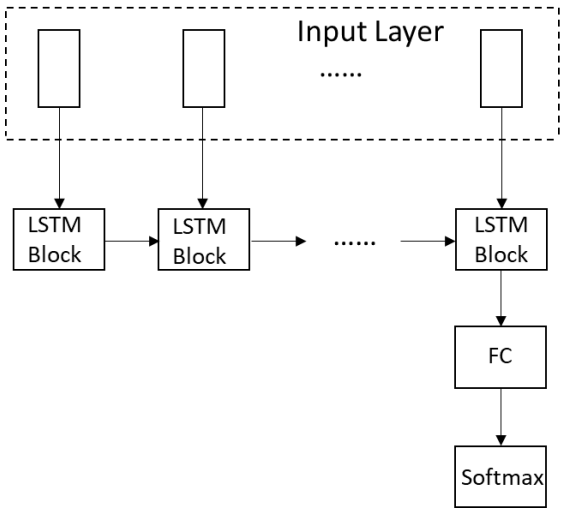


Fig.1 LSTM structure for human behavior recognition

In this paper, Kinect V2 equipment is used to obtain the positions of 25 skeleton points of the human body [7]. The human skeleton model is shown in Figure 2.

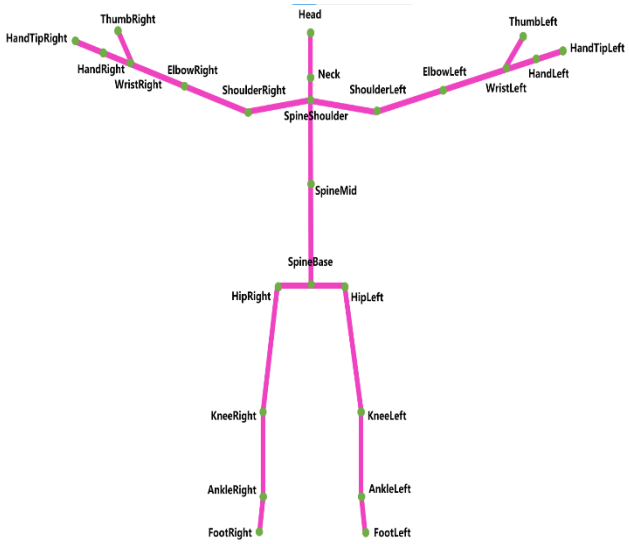


Fig.2 Human skeleton model

A picture of human skeleton obtained by Kinect is shown in Figure 3.

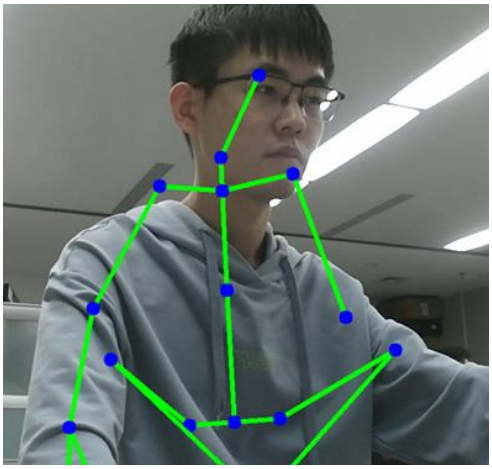


Fig.3 picture of human skeleton

In the experiment, a homemade dataset is used, including 7 action categories and 1310 samples. The labels for different actions are shown in Table 1.

Table.1 Labels of the homemade dataset

Label	0	1	2	3	4	5	6
Action	Walk	Jog	Sit	Handwave	Kick	Fall	Punch

2.2. Speech Emotion Recognition based on CNN

At present, the acoustic features used in the research of speech emotion recognition can be roughly divided into three categories: prosody features, sound quality features and spectral correlation features [8]. Spectral correlation features are used to analyze the sound in the frequency domain, which can better represent the human perception of the sound. The application of this kind of features to speech emotion recognition has achieved outstanding results in recognition accuracy and recognition effect.

Among the spectral correlation features, Mel frequency cepstrum coefficient (MFCC) is more prominent. It is based on human auditory perception, and its performance in emotional characteristics is closer to human body, which can effectively improve the accuracy of recognition. Therefore, this paper will extract MFCC features of sound signals for speech emotion recognition.

To solve the problem of speech emotion recognition, CASIA speech dataset and CNN model are used. In the CASIA speech dataset, emotions include the following six categories: Angry, Fear, Happy, Neutral, Sad, Surprise. The CNN network structure built in this paper is shown in Figure 4.

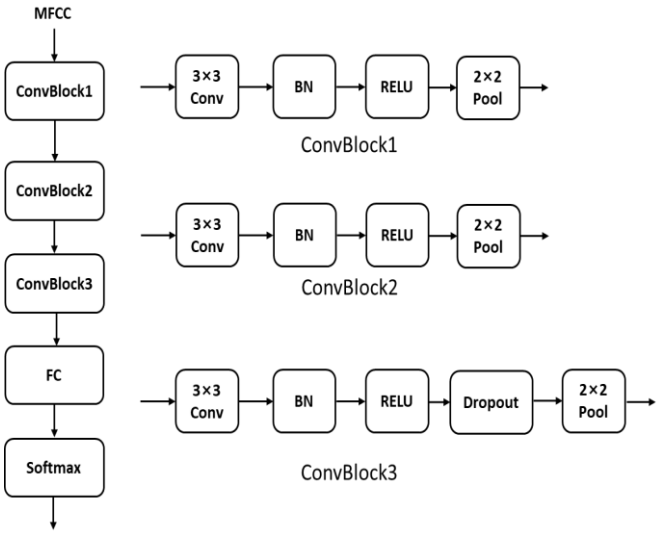


Fig.4 CNN network structure

3. MULTIMODAL EMOTIONAL ATMOSPHERE FIELD ESTIMATION METHOD

After classifying human behavior and speech emotion respectively, decision-making fusion method is needed. In this paper, the confidence of each classification result is calculated by a maximum weight multimodal information fusion algorithm, and then the danger degree of different classifications is assigned. Finally, the criteria for judging emotional atmosphere are established and the possible dangerous behaviors are warned and alarmed.

3.1. Maximum Weight Multimodal Information Fusion Algorithm

In decision-making fusion, it is necessary to count the classification results of each mode, establish certain criteria to calculate the classification results, and finally output the results. In view of the different recall rates for different classifications in each mode, this paper proposes a maximum weight multimodal information fusion algorithm to calculate the confidence of each classification result, which reduces the impact of different recalls to a certain extent. The specific implementation steps are as follows.

i. For all two modes, each mode has m classification results

ii. The recall of using the ith mode (i=1,2) to identify the jth result (j= 1,… , m) is p_{ij} , so the weight of ith mode $w_i = \begin{pmatrix} p_{i1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & p_{im} \end{pmatrix}$

iii. The probability of identifying the jth classification result on the ith mode at a certain time is c_{ij} , then the output of the ith mode through the softmax layer at that time $c_i = (c_{i1}, c_{i2}, \ldots, c_{im})^T$

iv. Confidence of all m results in the ith mode $C_i = w_i c_i = \begin{pmatrix} p_{i1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & p_{im} \end{pmatrix} c_i$

Through the above maximum weight multimodal information fusion algorithm, the confidence of each classification result in the two modes is obtained, where $C_1 = (p_{11}c_{11}, p_{12}c_{12}, p_{13}c_{13}, p_{14}c_{14}, p_{15}c_{15}, p_{16}c_{16}, p_{17}c_{17})^T$ $C_2 = (p_{21}c_{21}, p_{22}c_{22}, p_{23}c_{23}, p_{24}c_{24}, p_{25}c_{25}, p_{26}c_{26})^T$.

3.2. Criteria for Judging the Emotional Atmosphere

Before multimodal fusion, it is necessary to determine the risk degree of different classification results in each mode. Therefore, according to the risk levels of different human behaviors and speech emotions in public places, this paper assigns the risk level to each classification result, and the results are shown in Table 2 and Table 3.

Table.2 Human behavior assignment

Assignment	Action
0	Walk, Sit
1	Handwave
2	Jog, Fall
3	Kick, Punch

Table.3 Speech emotion assignment

Assignment	Emotion
0	Neutral, Happy, Surprise
1	Sad
2	Fear
3	Angry

According to the danger degree assignment in different modes and the order of output labels, the danger degree assignment matrix for the two modes are obtained, where $d_1 = (0,2,0,1,3,2,3)$, $d_2 = (3,2,0,0,1,0)$.

Since human behavior and speech emotion are equally important and indispensable for judging dangerous behaviors in public places, the two modes are given the same weight, i.e $\alpha_1 = 0.5, \alpha_2 = 0.5$. With the confidence, danger degree assignments and fusion weight of each mode, the weighted fusion rule is designed as shown in formula (1).

$$fv = \frac{d_1 C_1}{\sum_{j=1}^{7} C_{1j}} \times \alpha_1 + \frac{d_2 C_2}{\sum_{j=1}^{6} C_{2j}} \times \alpha_2 . \quad (1)$$

The result fv combines human behavior and speech emotion information, and represent the degree of danger in public places. According to the result fv , the emotional atmosphere of the crowd in public places is divided into three states: quiet, tense and excited. The quiet state represents a low degree of danger, the tense state represents a higher degree of danger, and the excited state represents an extremely high degree of danger. The classification criteria are as follows.

- If $fv < 1$, the atmosphere is quiet.
- If $1 < fv < 2$, the atmosphere is tense.
- If $fv > 2$, the atmosphere is excited.

When the crowd is in a quiet atmosphere, there is no need for early warning and alarm. When the crowd is in a tense atmosphere, give early warning of possible dangerous behaviors. When the crowd is in an excited atmosphere, it is likely that dangerous behaviors such as quarrels or fights are taking place in public places at this time, and the alarm will be given for this emergency

4. EXPERIMENT

In order to verify the effectiveness of the dangerous behavior recognition system established in this paper, this section will conduct experiments on human behavior recognition and speech emotion recognition, and finally give some examples to demonstrate the effect of multimodal fusion.

4.1. Experiment of Human Behavior Recognition

The experiment of human behavior recognition is to predict possible dangerous behaviors by analyzing human actions. The experiment is conducted on the basis of homemade behavior recognition dataset, which is collected in the classroom including 7 action categories. Taking 70% of the homemade dataset as the training set and 30% as the test set, the confusion matrix is obtained by training through the LSTM network, which is shown in Table 4.

Table.4 Confusion matrix of behavior recognition

	Walk	Jog	Sit	Handwave	Kick	Fall	Punch
Walk	49	3	0	0	1	0	0
Jog	11	35	0	0	4	0	2
Sit	0	0	61	0	0	2	0
Handwave	2	4	0	52	1	1	0
Kick	2	1	0	0	49	0	0
Fall	0	0	0	0	0	50	0
Punch	1	3	0	0	1	0	58

The confusion matrix shows the results on the test set of the homemade dataset, in which the left side is the real label and the upper side is the test label. The experimental results show that the accuracy of the model is 90%. For the dangerous actions, such as ‘Kick’, ’Fall’ and ‘Punch’, the precision of the three is 94.2%, 94.33% , 96.67%, the recall of the three is 94.2%, 100%, 92.1%. The result of the model is satisfied, so the next experiment can be carried out.

4.2. Experiment of Speech Emotion Recognition

The experiment of speech emotion recognition is to predict possible dangerous behaviors by analyzing human emotion. The experiment is conducted on the basis of CASIA speech dataset through CNN network. Take 70% of the CASIA dataset as the training set and 30% as the test set, and the confusion matrix is shown in Table 5.

Table.5 Confusion matrix of emotion recognition

	Angry	Fear	Happy	Neutral	Sad	Surprise
Angry	52	1	7	1	0	9
Fear	0	30	0	2	10	0
Happy	4	0	46	3	1	4
Neutral	0	3	3	54	3	0
Sad	0	13	0	1	40	0
Surprise	10	6	3	0	1	53

The confusion matrix shows the results on the test set of the CASIA dataset, in which the left side is the real label and the upper side is the test label. The experimental results show that the accuracy of the model is 76%. For the dangerous emotion ‘Angry’, the

precision is 78.79%, the recall is 79.31%. The result of the model is satisfied, so the experiment of multimodal fusion can be carried out.

4.3. Demonstration of Multimodal Fusion

For the multimodal fusion method proposed above, this section will give several examples to demonstrate the effectiveness of this method.

Among the human behavior mode, the recall of seven behaviors are 92.5%, 67.3%, 96.8%, 86.7%, 94.2%, 100% and 92.1%, so w_1 has been determined.

Among the speech emotion mode, the recall of six emotions are 74.29%, 71.43%, 79.31%, 85.71%, 74.07% and 72.6%, so w_2 has been determined.

Demonstrations are carried out when people are chatting, quarrelling and fighting, and the emotional atmosphere of the crowd was judged by fv.

The results when people are chatting are shown in Table 6.

Table.6 Result when people are chatting

	Classification	Output of Softmax	Confidence	Single result
Human Behavior	*Walk*	68.14%	63.03%	0.65
	Jog	7.96%	5.36%	
	Sit	1.77%	1.71%	
	Handwave	5.31%	4.60%	
	Kick	5.31%	5.00%	
	Fall	0.88%	0.88%	
	Punch	10.62%	8.61%	
Speech Emotion	*Angry*	5.76%	4.28%	0.29
	Fear	7.22%	5.16%	
	Happy	1.51%	1.20%	
	Neutral	82.13%	70.39%	
	Sad	1.33%	0.99%	
	Surprise	2.05%	1.49%	
Fusion result	0.47			

The result shows that the possibility of 'Walk' and 'Neutral' is the greatest in this situation, and there may be a chatting. The fusion result $fv = 0.47 < 1$, so the emotional atmosphere is 'quiet'.

The results when people are fighting are shown in Table 7.

Table.7 Result when people are fighting

	Classification	Output of Softmax	Confidence	Single result
Human Behavior	*Walk*	2.13%	1.97%	2.84
	Jog	1.34%	0.90%	
	Sit	0.88%	0.85%	
	Handwave	1.02%	0.88%	
	Kick	9.885	9.31%	
	Fall	3.54%	3.54%	
	Punch	81.21%	74.49%	
Speech Emotion	*Angry*	75.21%	55.87%	2.39
	Fear	3.59%	2.56%	
	Happy	1.55%	1.23%	
	Neutral	7.32%	6.27%	
	Sad	9.31%	6.90%	
	Surprise	3.02%	2.19%	
Fusion result	2.615			

The result shows that the possibility of 'Punch' and 'Angry' is the greatest in this situation, and there may be a fight. The fusion result $fv = 2.615 > 2$, so the emotional atmosphere is 'excited'.

Through the above results, it can be seen that the dangerous behavior identification method can accurately identify the dangerous behavior in reality, which has great application value. However, this section is only a demonstration of several examples, which needs to be verified in practice.

5. SUMMARY

Dangerous behaviors in public places have brought a great negative impact on the society, and there are many problems through manual identification of dangerous behaviors. In order to automatically identify dangerous behaviors in public places, this paper proposes a dangerous behavior identification system based on multimodal fusion. The main research of this system is divided into three aspects: one is the identification of human behavior which mainly analyzes dangerous actions, the second is speech emotion recognition which extracts the MFCC features and analyzes the emotions, the third is the multimodal fusion method which judges the emotional atmosphere of the crowd through the weighted fusion method. The actual test results show that the dangerous behavior identification system established in this paper can accurately identify dangerous behaviors in public places, which can be applied under some circumstances. In addition, the background of datasets

used in this paper is relatively single, so the dangerous behavior recognition of the model in the complex background will be the main research direction in the future.

Acknowledgements

This work is supported by the Beijing Municipal Natural Science Foundation under Grant L191020 and Systematic Key Research P2021T002 of China State Railway Group Co., Ltd.

REFERENCES:

[1] Johansson Gunnar, Visual perception of biological motion and a model for its analysis. Perception & Psychophysics,1973.

[2] Han Jungong,Shao Ling,Xu Dong,Shotton Jamie, Enhanced computer vision with Microsoft Kinect sensor: a review. IEEE transactions on cybernetics,2013.

[3] Wu Zuo, Human Motion Tracking Algorithm Based on Image Segmentation Algorithm and Kinect Depth Information. Mathematical Problems in Engineering,2021.

[4] Ghosh S,Laksana E,Morency LP,Scherer S, Representation Learning for Speech Emotion Recognition. InInterspeech 2016, pp. 3603~3607.

[5] Lina,Arlends Chris, Recognizing Human Emotions using Emotional Transition Lines in Eigenspace. Proceedings of 2010 2nd International Conference on Multimedia and Computational Intelligence (ICMCI 2010).Institute of Electrical and Electronics Engineers,2010,pp. 316-319.

[6] Sreevidya P,Veni S,Ramana Murthy O V. Elder emotion classification through multimodal fusion of intermediate layers and cross-modal transfer learning. Signal, image and video processing,2022(prepublish).

[7] Bhateja, Aditi,Shrivastav, Adarsh,Chaudhary, Himanshu,Lall, Brejesh,Kalra, Prem K, Depth analysis of kinect v2 sensor in different mediums. Multimedia Tools and Applications,2021, in press.

[8] Tian Feng,Shuying Yang, Speech Emotion Recognition Based on LSTM and Mel Scale Wavelet Packet Decomposition. Proceedings of 2018 International Conference on Algorithms, Computing and Artificial Intelligence (ACAI 2018).,2018,pp. 230-236.