# An Emotion Recognition System Based on Human Behavior in Passenger Transport

**Qiwei Yu[1], Yaping Dai[1], Kaoru Hirota[1] ,Zhiyang Jia[1],and Wei Dai[2]**

1.School of Automation, Beijing Institute of Technology, Beijing 100081, China
E-mail: yqw2000@foxmail.com
2.River Security Technology Co., LTD, Shanghai 200336, China

**Passengers' abnormal emotions such as nervous and anger often lead to traffic accidents in passenger transport. Surveillance of abnormal emotions possesses significant potential for increased security within passenger transport. An emotion recognition system based on human behavior is designed in this paper. Three emotions of quiet, nervous, and anger are set in research. The system uses Long short-term memory to process human behavior data, and recognizes whether passengers have abnormal emotions such as nervous or anger based on behavioral analysis. Experiments on the recognition performance of the system, the experimental results show that the system has an accuracy of more than 95%, and the recognition period is between 2 to 3 seconds.**

**Keywords: Emotion Recognition, Body Behavior, Recurrent Neural Networks, Passenger Transport Safety**

## 1. INTRODUCTION

Many traffic accidents are caused by abnormal emotions of passengers, such as nervous and anger. Therefore, identifying whether abnormal emotions occur has an important impact on traffic safety[1]. Schuller et al. introduced audiovisual recognition of emotion and behavior for surveillance in public transport systems[2]. Zhang et al. studied the intonation and speed of speech signal to recognize three emotions, which are “Neutrality”, “Happiness” and “Anger”[3]. To effectively recognize driver's emotions, t Du et al. proposed a new deep learning framework called Convolution Bidirectional Long Short-term Memory Neural Network (CBLNN). CBLNN uses Multi-modal factorized bilinear pooling (MFB) to fuse the extracted information and classifies it into five common emotions: happiness, anger, sadness, fear and neutrality[4]. Most of these methods recognize emotions by analyzing data such as facial expressions or speech. However, the collection of facial expressions and voices is easily disturbed by noise in some special situations. For example, during the epidemic, people are often required to wear masks when taking transportation, which makes it difficult to obtain effective facial data.

Therefore, an emotion recognition system is proposed for the passenger transport, which analyzes the behavior of the human body to obtain the emotional state of people, and promptly warns when abnormal emotions appear. The passenger carriage environment is simulated indoors in this research, and emotions are divided into three categories: quiet, nervous, and anger. Quiet is normal emotion, and nervous and anger are abnormal emotions. The emotion recognition system uses long and short-term memory networks to process behavioral data, and extract the human behavior information of multiple people at the same time and recognize emotions. Compared with the research on emotion recognition based on face and voice, the emotion recognition system based on human behavior can perform better in special environments such as the face is blocked or the environmental noise is too large.

The proposed emotion recognition system mainly includes two parts, hardware module and software module, which work independently and communicate data through Socket. The hardware module obtains the human behavior joint data in the carriage and sends it to the software module, receives the emotion recognition result and displays the current human behavior skeleton image and the corresponding emotion category in real time; the software module processes the data and inputs the data to the neural network model, After getting the emotion recognition result, send it to the hardware module. Experiment with the emotion recognition system by changing the light, the number of people, and the data processing conditions. The test results show that the accuracy of the system can reach more than 95%, and the emotion recognition cycle is between 2 to 3 seconds (determined by the number of people in the scene). , which indicates that the system has a high recognition accuracy.

The design of the emotion recognition system is introduced in section 2. Experimental verification and result analysis are described in section 3.

## 2. DESIGN OF THE EMOTION RECOGNITION SYSTEM

In this section, The overall structural design of the emotion recognition system is introduced and the theoretical methods of the two modules are elaborated.

### 2.1. The Structure of the System

This work was supported by the Beijing Municipal Natural Science Foundation [grant number L191020].

The passenger carriage environment is simulated indoors. Passenger emotions are divided into three categories: quiet, nervous, and anger, each emotion has its own human behavior. The quiet behavior is sitting, the nervous behavior is standing up and sitting down repeatedly, and the anger behavior is punching. The emotion recognition system can recognize the human behavior of multiple people at the same time, and process and classify the behavior data to obtain the emotion of the person.

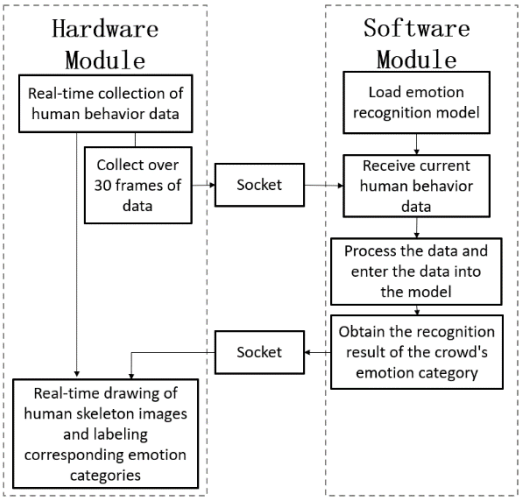


**Fig. 1** The structure of the emotion recognition system.

The overall workflow of the emotion recognition system is shown in Fig. 1. The system is divided into two parts: the hardware module and the software module. The two parts operate independently and carry out data communication through Socket. The hardware module uses Kinect to obtain the human body behavior data in the carriage. When the number of frames is collected to the specified number of frames (30 frames), the data is packaged and transmitted to the software module; after the software module receives the data, it processes the data and enters the emotion recognition model, and then the recognition result is passed to the hardware module; after the emotion recognition result is obtained, the hardware module displays the current human skeleton image and the corresponding emotion category in the screen in real time.

### 2.2. Hardware module

The hardware module contains two functions: real-time acquisition of human behavior joints in the scene and sending to the software module; receiving emotion recognition results, drawing human skeleton images on the screen and text annotations next to the human body.

When people make different behaviors, the corresponding joint points have different position information and angle information. Therefore, tracking the coordinates of the bone joint points becomes a direct and reliable method to realize behavior recognition. For a human behavior, the content of the behavior can be effectively expressed through multiple human joints.

The hardware module obtains human behavior information through Kinect.Kinect detect people within 0.5 to 4.5 meters in the scene, and can recognize the bone data of up to 6 people at the same time[5]. The human body joint point data tracked by Kinect is represented by the X, Y, Z three-dimensional coordinates in the Kinect camera coordinate system.

Human behavior can be decomposed into time series related to body shape. The body shape is represented by 3D joint points of various parts of the human body. The data sample of each behavior can be modeled as a sequential representation of multiple frames of bone data. The data of each human behavior is composed of 30 frames of bone information in this study, and each frame of information contains the (x, y, z) three-dimensional coordinates of 25 sets of joints of the human body[6].

Holt two-parameter linear exponential smoothing method can smooth the joint data of human body behavior and reduce jitter. In the emotion recognition system, users can choose whether to use holt's two-parameter linear exponential smoothing method to process data. Holt two-parameter linear exponential smoothing mainly includes two smoothing formulas. The two smoothing formulas smooth two factors of time series respectively as shown in (1) and (2).

$$S_t = \alpha x_t + (1-\alpha)(S_{t-1} + b_{t-1}) \tag{1}$$

$$b_t = \gamma (S_t - S_{t-1}) + (1-\gamma) b_{t-1} \tag{2}$$

$\alpha$ and $\gamma$ are smoothing parameters; $x_t$ is the actual observation value; $b_t$ is the trend value, $S_t$ is the smoothing value.

### 2.3. Software module

The software module processes the received human behavior data, inputs the data into the trained emotion recognition model, and returns the recognition results obtained to the hardware module.

The software module uses coordinate transformation to convert the coordinates of the joint points of the human body from absolute coordinates to relative coordinates, thereby reducing the impact of the location of the tester. Among the 25 sets of joint points detected by kinect, the joint with serial number 0 corresponds to the center of the human body. Select the joint point with serial number 0 as the reference point, subtract the coordinates of the reference point from the joint point coordinates of the other 24 groups of body parts, and translate the joint coordinates of the entire person to the coordinate system with the center of the human body as the origin. The data information between the joints is unchanged, but the relative coordinate difference of the human joint points under different behaviors is highlighted.

The emotion recognition model is built by the neural network, and the structure is shown in Fig. 2.

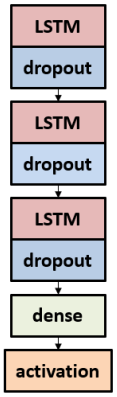


**Fig. 2** Structure of emotion recognition model.

Long Short-Term Memory (LSTM) network is a specially improved network based on RNN, which can learn long-term dependent information[7]. LSTM maintains long-term memory in each RNN unit and learns when to remember or forget the information stored in its internal memory unit. The three LSTM layers are stacked together, so that the model can have a higher level of temporal representation capabilities. The dimensions of the output space of the three LSTM network layers are all set to 5. The first two layers of LSTM network return the complete output sequence; the third layer LSTM network only returns the last step of the output sequence, converting the input sequence into a single vector, reducing the time dimension.

In order to improve the over-fitting problem, each layer of the LSTM network is equipped with a Dropout layer for random parameter discarding, and the ratio of the Dropout layer is set to 0.5[8]. After the model passes through the LSTM layer and the Dropout layer, the data is input to the fully connected layer Dense. The dimension of the output space of the Dense layer is set to the number of categories of the training dataset; the last layer of the model is the activation function Softmax.

### 2.4. Data Communication

The hardware module and the software module operate independently, and transfer data communication through Socket[9]. Socket is a communication mechanism in network programming, and is a software abstraction layer for communication between the application layer and the TCP/IP protocol suite[10]. Socket is a set of interfaces used to describe IP addresses and ports. It hides the complex TCP/IP protocol family. Users only need to face a set of interfaces and use Socket to organize relevant data to conform to the specified protocol.

Socket communication created by default is blocked. Blocking mode means that the current thread is suspended until the call result is returned. The non-blocking mode is the opposite of the blocking mode. Even if the process cannot get the result immediately, the call will not block the current thread. In the emotion recognition system, configure the Socket of the hardware module to non-blocking mode. When the Socket is performing an operation, even if there is no data in the receiving buffer at this time, the thread will not be allowed to wait. No matter whether the operation is completed or not, the called function will immediately return to the working mode. Set the Socket receiving of the software module to blocking mode, and the software module will work only when the hardware module sends data information.

## 3. EXPERIMENTS AND ANALYSIS

The experiment and result analysis of the emotion recognition system are introduced in this section.

### 3.1. Experimental Verification

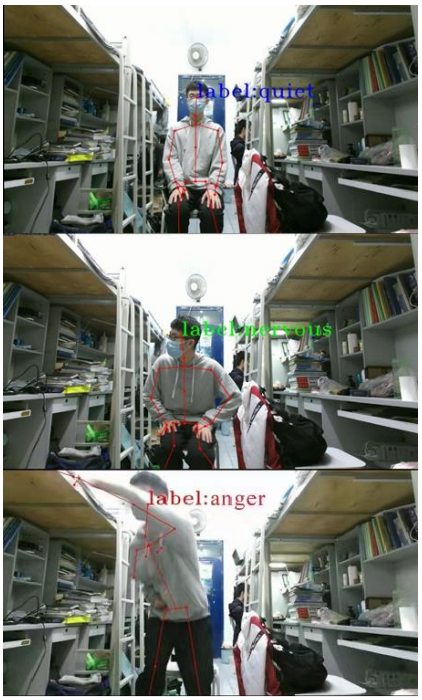


**Fig. 3** Experimental results.

The experimenter completed the three actions of sitting, standing up and sitting repeatedly, and punching. The emotional category of sitting is quiet, and the text label is "quiet"; the emotional category of standing up and sitting repeatedly is nervous, and the text label is "nervous"; The emotional category of punching is anger, and the text label is "anger". The image result output by the hardware module in the experiment is shown in Fig. 3.

Change the number of people, light, data processing and other conditions respectively for comparison, and conduct experiments on the performance of the emotion recognition system under different conditions. The data results measured in the experiment are shown in Table 1.

### 3.2. Result Analysis

From the experimental results, it can be concluded that the overall accuracy of the emotion recognition system is more than 95%, and abnormal emotions in the human body can be recognized in a timely and accurate manner.

**Table. 1** Experimental verificatiol.

| experiment | Testing Under Different Environmental Factors | | | | |
|---|---|---|---|---|---|
| | ***Light*** | ***people*** | ***Data processing*** | ***Accuracy*** | ***Update delay*** |
| 1 | Bright | One | Not | 99.4%——99.8% | 2.29s——2.43s |
| 2 | Bright | Two | Not | 95.2%——99.8% | 2.31s——2.75s |
| 3 | Dusky | One | Not | 99.3%——99.8% | 2.31s——2.40s |
| 4 | Bright | One | Holt two-parameter linear exponential smoothing | 99.6%——99.9% | 1.98s——2.04s |

In this experiment, because there are only three emotions and the data scale is small, the accuracy of the model recognition is high.

Increase the number of people and compare the results of experiment 1 and experiment 2. As the number of people increases, the accuracy of system identification will decrease slightly. In the case of a single person, the delay time of the recognition result of the emotion category is between 2 to 2. 4 seconds. Every time the number of people in the scene increases by one person, the recognition result of the emotion category will increase the delay time between 100 to 300 milliseconds.

Changing the light factor and comparing the results of experiment 1 and experiment 3, it can be seen that the emotion recognition system that uses Kinect to obtain behavioral data is less affected by light, and can recognize emotions even in dark conditions.

In the comparative experiment 4 of data smoothing, because the bone movement processed by holt's two-parameter linear exponential smoothing method is smoother and inconsistent with the original human movement, the bones are presented separately in experiment 4. Since there is no need to convert RGB coordinates, and there is less noise in the data, the update delay of the emotion category is slightly faster.

## 4. CONCLUSION

An emotion recognition system applied to the carriage environment in passenger transport is designed in this paper, which can recognize emotions through the behavior data information of human body in the carriage. The carriage environment in passenger transport is simulated, and the emotion recognition model based on LSTM is used to process human behavior data. People's emotions are divided into three types: quiet, nervous or anger, which could be displayed in time to inform the staff when passengers had abnormal emotions. It has high practicability and effectively reduced potential safety hazards. Change the light, the number of people, and the data processing conditions to recognize the three emotions. The results show that the accuracy of the system can reach more than 95%, and the emotion recognition cycle is between 2 to 3 seconds (determined by the number of people in the scene). which shows that the emotion recognition system has great recognition accuracy and robustness.

In the future research, we will increase the variety and scale of human behavior data for research, improve the accuracy and generalization ability of the system, and better serve the passenger transport safety.

## REFERENCES:


[1] C. D. Katsis, G. Rigas, Y. Goletsis, and D. I. Fotiadis, "Emotion Recognition in Car Industry," in Emotion Recognition, Hoboken, NJ, USA: John Wiley & Sons, Inc., 2015, pp. 515–544

[2] B. Schuller, M. Wimmer, D. Arsic, T. Moosmayr, and G. Rigoll, "Detection of security related affect and behaviour in passenger transport, " Interspeech, Conference of the International Speech Communication Association, Brisbane, Australia, September. DBLP,2008.

[3] Y. Zhang, Y. Chen, and C. Gao, "An effective deep learning approach for dialogue emotion recognition in car-hailing platform," in 2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), 2019.

[4] G. Du, Z. Wang, B. Gao, S. Mumtaz, K. M. Abualnaja, and C. Du, "A convolution bidirectional long short-term memory neural network for driver emotion recognition," IEEE Trans. Intell. Transp. Syst., vol. 22, no. 7, pp. 4570–4578, 2021.

[5] J. Shotton et al., "Real-time human pose recognition in parts from single depth images," Commun. ACM, vol. 56, no. 1, pp. 116–124, 2013.

[6] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016

[7] C. Ye, X. Li, and J. Chen, "A deep network for tissue microstructure estimation using modified LSTM units," Med. Image Anal., vol. 55, pp. 49–64, 2019.

[8] J.-Y. Jiang and C.-T. Li, "Forecasting Geo-sensor data with participatory sensing based on dropout neural network," in Proceedings of the 25th ACM International on Conference on Information and Knowledge Management - CIKM '16, 2016.

[9] L. Hui-ping, C. Dai-min, and Y. Miao, "Communication of multi-robot system on the TCP/IP," in 2011 International Conference on Mechatronic Science, Electric Engineering and Computer (MEC), 2011.

[10] "Introduction to TCP/IP," in CompTIA A+ Complete Study Guide, Hoboken, NJ, USA: John Wiley & Sons, Inc., 2019, pp. 363–402.