Indoor Crowd Posture Recognition and Emotion Perception

Yishan Chen[1,2], Yaping Dai[1,2], Kaoru Hirota[1,2], Zhiyang Jia *[1,2]

*[1] School of Automation, Beijing Institute of Technology, Beijing, 100081, China
E-mail: yishan_chen@bit.edu.cn
*[2]State Key Laboratory of Intelligent Control and Decision of Complex Systems, Beijing, 100081, China
E-mail: yishan_chen@bit.edu.cn

Abstract. In order to detect the abnormal human beings' emotions in public hall, an emotional awareness alarm (E-alarm) system is proposed to perceive crowd emotions. The E-alarm system should monitor the indoor crowd posture by image analysis of monitoring video and identifies three emotions of "calm", "nervous" and "angry". The alarm emails will be sent automatically when getting "nervous" and "angry" emotions. The system is required to achieve rapid alarm adapted to the complex indoor environments in a short working cycle with high accuracy. It uses Microsoft Kinect to extract typical posture features precisely and machine learning models to achieve the stable performance in the continuous working period. Experimental result shows that E-alarm system can effectively identify human emotions and alarm abnormal emotions. The alarm cycle is about 30s, and the alarm accuracy rate reaches 91.7%.

Keywords: Human Posture Recognition, Emotion perception, Convolutional Neural Network, Kinect Sensor

1. INTRODUCTION

Recent studies of human emotion recognition have become one of the mainstream fields, and the emotion recognition in human-computer interaction has been widely concerned. Emotion recognition has made great contributions to human-computer interaction. V. Gentile et al. connected the relationship between body movements and spoken user sentences to reveal user's emotion from gesture [1]. T. M. W. Vithanawasam presented an approach to recognize the face and upper-body emotions for service robots by using linear discriminant analysis and pattern recognition neural network classifiers [2]. Sanket Agrawal et al. designed a system to extract facial expression features from Cohn-Kanade database and identified emotions by a ResNet-18 based Convolution Neural Network (CNN) model. The system could be used to find the level of interest that customers have in products displayed on a particular shop window [3]. Skeleton modeling and facial recognition based on deep learning can improve the accuracy of emotion recognition. H. Zhang acquire head and shoulder movements to classify the movements of the body when amusement is evoked by using decision tree, random forest, XGBoost, Support Vector Machine (SVM), and neural network [4]. Patwardhan et al. extracted the facial expression, head, hand and body movement and speech tracking for detecting anger and aggressive actions. Recognition was achieved using SVM and rule-based features, which improved affect recognition precision rate for anger [5]. Y. Maret et al. proposed a real-time system capable of recognizing four gestures by using SVM that correlate to human emotions based on the arm movements. The confidence of SVM was above 85% for each gesture [6]. Ferdous Ahmed made a unique combination of Analysis of Variance (ANOVA) and Multivariate Analysis of Variance (MANOVA) to eliminate irrelevant features. Besides, a binary chromosome-based genetic algorithm is proposed to maximizes the emotion recognition rate. The system could achieve recognition accuracy of 86.66% in an action-independent scenario [7]. However, previous research focused only on a limited number of movement features from a vast number of computable features and cannot adapt to the complex environment in large public places. It is required to design a system with both real-time and accuracy.

This paper proposes an E-alarm system to collect and monitor the crowd posture in real-time through image recognition and machine learning technology. The system obtains the features from human beings' posture, and divides the crowd emotions into three categories: "calm", "nervous" and "angry". Telnet technology is used to send expected emails continuously in different real-time emotions. The working period is about half a minute and enables users to make response to the emergencies in time. The system consists of three modules: posture collection module, emotion perception

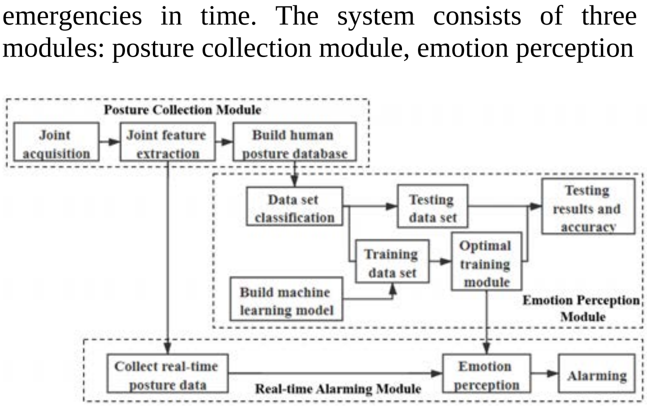
Posture Collection Module
Joint acquisition
Joint feature extraction
Build human posture database
Data set classification
Testing data set
Testing results and accuracy
Training data set
Optimal training module
Build machine learning model
Emotion Perception Module
Collect real-time posture data
Real-time Alarming Module
Emotion perception
Alarming

Fig. 1 The overall structure of the project.

module, and real-time alarming module. The overall structure is shown in **Fig. 1**.

The posture collection module uses Microsoft Kinect to collect human images. The posture database, contains the coordinates of 25 joints collected by Kinect sensor, will be established after data preprocessing. The emotion perception module uses Fast Fourier Transform (FFT) to extract the frequency domain features. Besides, two machine learning models, CNN and SVM, are used to analyze the indoor emotion. The optimal training module, evaluated by the accuracy of the testing results, will be selected as the criterion of emotion perception. The real-time alarming module uses telnet technology to send alarm emails according to the perception results, so as to achieve the rapid alarming in emergency.

The remainder of this paper is organized as follows. The scheme of obtaining posture features by Kinect and the steps of data preprocessing is introduced in 2. The optimal module of emotion perception is established by CNN and SVM in 3. The establishment of E-alarm system is introduced in 4. The system performance is verified in 5. This paper ends with the conclusion and the future work introduction in 6.

2. POSTURE DATA ACQUISITION AND PREPROCESSING

2.1. Data acquisition

Due to the complexity of the environment in large public places and the variability of crowd posture, using RGB images cannot solve the problem of human overlapping. With the wide application of Microsoft Kinect, its function of multi-person recognition and depth image brings more possibilities in the field of emotion perception. Therefore, Kinect will be selected for human posture recognition based on joint tracking.

The Kinect sensor device can track the movements associated with the user's 25 joints and collect information at the rate of 30 frames per second for less than 10 people at the same time. Each joint has its corresponding number tag recorded as 0 to 24. The joints obtained by Kinect v2 is shown in **Fig. 2**. The posture collection platform will collect 25 joint data in the 3D space and represent data in a depth vector (x, y, z).

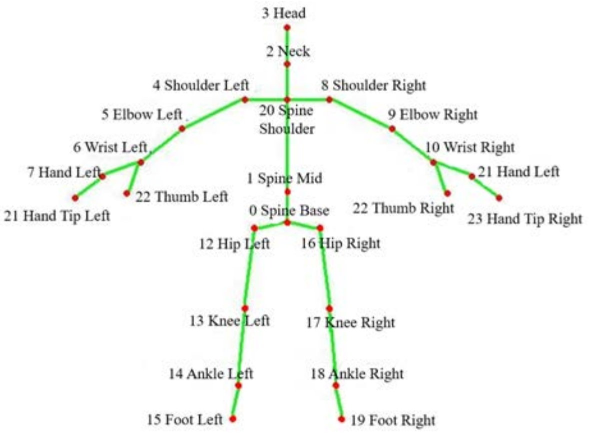

Fig. 2 Joints obtained by Kinect v2.

Two kinds of representation methods, joint position representation and joint angle representation, are used to record the posture features of the collected data. Joint position representation integrates the three-dimensional coordinates of each original joint into a vector. Each vector contains 75 coordinate data of 25 nodes in the 3D space. The vectors collected in continuous time can be combined into a set of time-space vectors. However, the disadvantage is that the difference in human figures will result to the accumulation of bias on all joints. The distance between Kinect and participants also causes error. The collected vector is shown in Eq. (1).

$$f_{JP}=\{C_i \mid i=0,1,2,\cdots\cdots,74\} \tag{1}$$

Joint angle representation is used to describe posture features by calculating the angle between each joint. It can not only concentrate on the bias caused by part of the joint movements, but also avoid the influence of human figure. Furthermore, it eliminates the difference between human body coordinate frame and Kinect coordinate frame [8]. The steps of angle calculation are as follows.

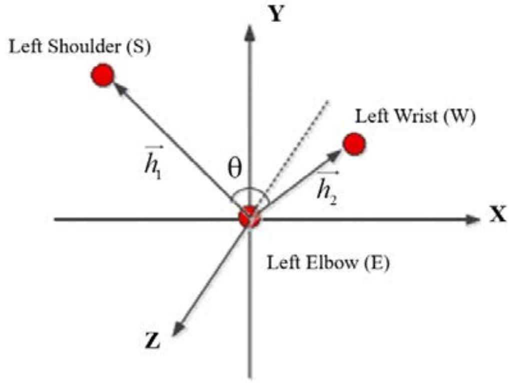

Fig. 3 Schematic diagram of joint angle calculation.

Firstly, calculate the vector of joints. As shown in **Fig. 3**, the three joints are left shoulder, left elbow and left wrist. Setting the coordinates of the left shoulder S as (S_x, S_y, S_z), the coordinates of the left elbow E are (E_x, E_y, E_z), and the coordinates of the left wrist W are (W_x, W_y, W_z). The vector from E to S is $\vec{h_1}$ and the vector from E to W is $\vec{h_2}$. The equations of $\vec{h_1}$ and $\vec{h_2}$ are shown as Eq. (2) and Eq. (3).

$$\vec{h_1}=(S_x-E_x, S_y-E_y, S_z-E_z) \tag{2}$$

$$\vec{h_2}=(W_x-E_x, W_y-E_y, W_z-E_z) \tag{3}$$

Secondly, the angle θ between two vectors can be calculated as shown in Eq. (4). The calculation of the angles of other joints are the same.

$$\cos\theta=\frac{\vec{h_1}\cdot\vec{h_2}}{|\vec{h_1}|\cdot|\vec{h_2}|} \tag{4}$$

There are 15 typical angles of human body extracted by joint angle representation method. The collected data includes position and angle can fully describe the posture features of the movements.

2.2. Data preprocessing

The original collected data should be preprocessed before establishing the database. Using data preprocessing procedure can effectively reduce the influence of noise and relieve the overfitting problem in the process of emotion classification. The procedure showed in **Fig. 4** includes three steps: Kalman Filtering, Coordinates Transformation and Coordinates Difference.

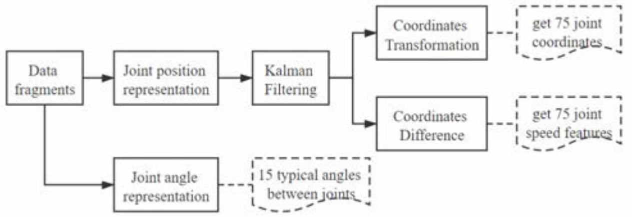
Data fragments
Joint position representation
Kalman Filtering
Coordinates Transformation
get 75 joint coordinates
Coordinates Difference
get 75 joint speed features
Joint angle representation
15 typical angles between joints

Fig. 4 The procedure of data preprocessing.

2.1.1. Kalman filter

The original data is mainly affected by two kinds of noise. One caused by disturbance and another caused by estimate error. In order to reduce the two kinds of noise, Kalman filter (KF) should be applied to data preprocessing.

Compared with the average filter, KF can obtain the optimal estimation of the posture state and avoid the fuzzy or loss of features. The coordinate data is taken as the input signal of KF, so as to obtain the output data with less affected by noise [9]. The relevant parameters, state matrix and observation matrix of KF are shown in the following formula.

We define the Kalman filtered state vector ${x_k}^T$ as a six-dimensional vector.

$$ {x_k}^T = \left(X, Y, Z, V_X, V_Y, V_Z\right) \tag{5} $$

The state transition matrix A is shown in Eq. (6).

$$ A = \begin{bmatrix} 1 & 0 & 0 & d_t & 0 & 0 \\ 0 & 1 & 0 & 0 & d_t & 0 \\ 0 & 0 & 1 & 0 & 0 & d_t \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{6} $$

The observation matrix H is shown in Eq. (7).

$$ H = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \tag{7} $$

As the observed measurement is $\left(X^{\text{Kinect}}, Y^{\text{Kinect}}, Z^{\text{Kinect}}\right)$, the estimated noise covariance matrix Q is defined in Eq. (8).

$$ Q = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{8} $$

The observation noise covariance matrix is defined in Eq. (9).

$$ R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{9} $$

After KF, the noise in the original data will be suppressed. Using filtered data for emotion perception will achieve higher accuracy than without KF.

2.2.2. Coordinates transformation and difference

As the collected data contains the displacement and the movements of body, it is difficult to concentrate on the moving process of joints instead of spatial position. The coordinates transformation method is shown in **Fig. 5.** It used to translate joints into a new frame with the origin becoming the base of spine instead of Kinect. Through this method, the system will pay more attention to the movement of human body rather than the displacement in space [10].

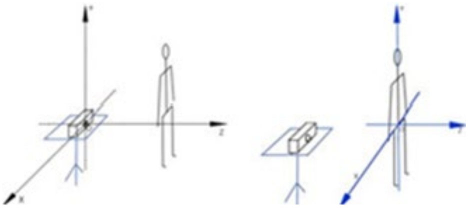

Fig. 5 The 3D coordinate system transformation.

In order to make the posture data attach to the time characteristics, the speeds of joints are added in the data preprocessing procedure. The joint velocity can be calculated by coordinates difference. For example, at time t, the system collects the coordinates x_t of the current frame and x_{t+1} of the next frame. The sampling cycle of Kinect is *T*. Then the velocity v_t of the joint at time t shows as Eq. (10). The calculation of the velocity of all joints in each coordinate component are the same.

$$ v_t = \frac{x_{t+1} - x_t}{T} \tag{10} $$

3. The establishment of emotion perception module

In this paper, CNN and SVM are used to realize the emotion perception. The optimal one will be selected to build the E-alarm system.

Before establishing the emotion perception module, posture data should be transformed in advance to extract the frequency domain features on the time series of each joint by using FFT. As an equivalent method of Discrete Fourier Transform (DFT), FFT uses more transforms with easier calculation and realizes large-scale transformation in a high speed. The amplitude and phase components of each joint obtained by FFT can be used as effective features and applied in machine learning. Besides, the posture data collected in different emotions should be labeled as 1 to 3. The label is simplified by the one-hot code.

The CNN model is improved from LeNet-5 to realize emotion perception with fast training speed [11]. The network structure is shown in **Fig. 6**. This model is composed of nine layers: an input layer, four convolutional layers, two maximum pooling layers, an average pooling layer, and a fully connection layer. The number of kernel functions of convolutional layer 1 are 16, the size of convolution window is 8, and the step size is 2. In the convolutional layer 2, the number of kernel functions are changed to 8, and the size of convolution window is changed to 4. Setting the parameters of the maximum pooling layers with the same step size of 2. We use Adam optimization algorithm to improve the training speed and avoid falling into local optimum. The 'Dropout' function is used to relieve overfitting problem. Different from the LeNet-5 model with single convolutional layer and single pooling layer, this model uses the optimization method of two identical convolutional layers in series. The LeNet-5 model only performs one nonlinear operation at a time, resulting in the insufficient extraction of features each time. The optimization method increases the number of non-linear operations and reduces the size of the convolution kernel. Therefore, it improves the training speed and optimizes the learning ability.

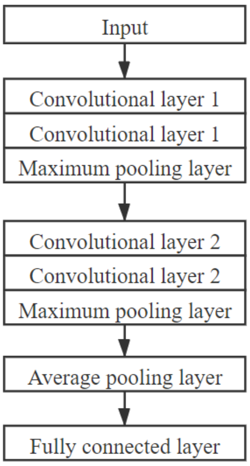

Fig. 6 The structure of CNN.

SVM is a kind of linear classifier which can be used for binary classification of data. In order to solve the three-classification problem of emotion perception, we use three SVM models to build a one-to-one combination scheme. The one-to-one combined classifier only needs $\frac{N(N-1)}{2}$ SVM models for N sample labels. Each SVM only classifies two kinds of tags recorded as 1 and -1. In the training process, each classifier outputs the classification results. The optimal class with the largest number of positive outputs will be selected as the final classification result [12].

4. THE ESTABLISHMENT OF E-ALARM SYSTEM

The E-alarm system includes three parts: the posture collection module, the emotion perception module, and the real-time alarming module. The structure of the system is shown in **Fig. 7**. The posture collection module can effectively monitor people below 10 in real time and record the coordinates, motion speed, and angle of the joints by using Kinect in real time. Then using data preprocessing to reduce the noise in the original data. The emotion perception module uses CNN to classify the emotion of the crowd from 30 consecutive frames in a period of time. The real-time alarming module uses SMTP and telnet protocol based on remote login technology to send e-mail to the target mailbox. It can read the emotion perception results and send corresponding e-mail when the emotion is nervous or angry. The e-mail includes the environmental image and the emotion of crowd.

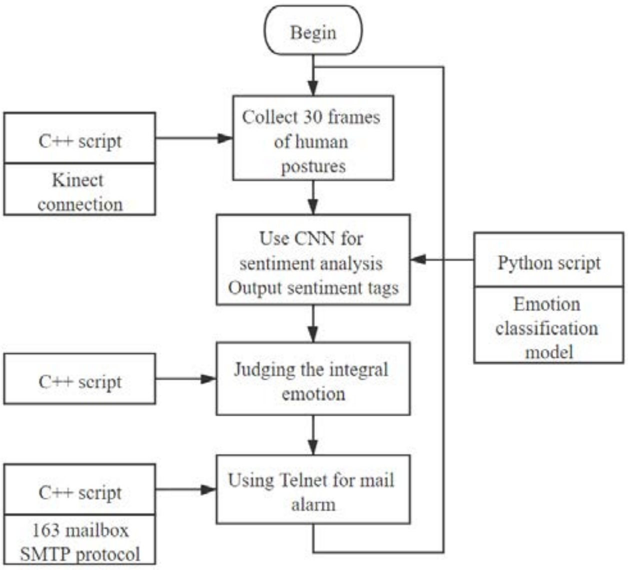
Begin
C++ script
Kinect connection
Collect 30 frames of human postures
Use CNN for sentiment analysis Output sentiment tags
Python script
Emotion classification model
C++ script
Judging the integral emotion
C++ script
163 mailbox SMTP protocol
Using Telnet for mail alarm

Fig. 7 The structure of the E-alarm System.

The E-alarm system adopts loose coupling mode of C++ script and python script. Using the control character to control the system. In this way, the system realizes data interaction between two scripts and avoids conflict. The loose coupling method not only ensures the real-time performance of the system, but also solves the problem that Microsoft Visual Studio cannot support TensorFlow library in Python script.

The control character is recorded in an independent file and will be rewritten accordingly when a module runs out. Two scripts query the control character at the same time in order to clarify which module should be run. When the control character is 1, the C++ script stops working and the python script starts to run. The emotion perception module analyzes the collected data and outputs the emotion categories of each frame. When the control character is 2, the python script stops running. Then the real-time alarming module of C + + script starts to work. When the control character is 3, the current cycle of E-alarm system has been completed, and the next cycle of data collection will start.

5. THE VERIFICATION OF E-ALARM SYSTEM

5.1. Experimental environment and preparation

The experiment carries out in the space of 4m*4m cross section. In order to pursue high-precision joint data, we need to ensure that the human body is not blocked by obstacles. The experimental environment is shown in **Fig. 8**.

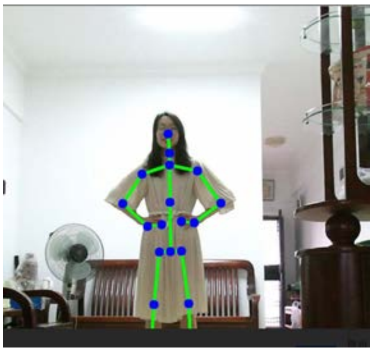

Fig. 8 Experimental environment.

There are four healthy residents take part in the experience. All of them have no illness that influence their motions. In order to ensure the validity of the experimental data, emotional guidance should be given to the participants before the experiment. During the experiment, the participants walk freely or spontaneously make body movements within a certain range of activities. At the same time, the posture collection module records the movement process of the participants within one minute. After the collection, the participants will describe the current emotional state again to ensure their emotions are the same before and after the experiment. There should be at least one day interval between each emotion feature collection experiment to avoid emotional interference.

5.2. Evaluation of Emotion Perception module

Human posture data collected in advance by using posture collection module should be divided into training sets and testing sets. Different emotions in the posture database were manually labeled to make 1200 training sets and 600 testing sets. Using CNN and SVM to load the training sets, and the one with high accuracy can be used as the criterion after training. Finally, the emotion perception module output the results of testing sets according to the criterion.

The working cycle of the testing module is 1s. The classification accuracy of CNN and SVM for different emotions is shown in **Table 1**. The testing module is established based on the optimal training module of CNN to classify 600 frames of testing data with 200 frames for each emotion. There are 523 correct classification results, and the accuracy rate reaches 87.2%. For SVM, it can get a good classification result after 1000 iterations and has a fast running speed. The testing module outputs 501 correct results with an accuracy of 83.5%. The classification accuracy of CNN is higher than that of SVM. The classification accuracy rate of calm and nervous is much higher than that of SVM, and the accuracy rate of angry is slightly lower. From the experimental results, we can draw the conclusion that the traditional classification pattern used by SVM performs slightly inferior to CNN in dealing with high-dimensional classification problems. Therefore, the E-alarm system finally uses CNN to identify emotions.

Table. 1 Classification results of Emotion Perception module

Emotion	SVM		CNN	
	Correct results	Accuracy	Correct results	Accuracy
Calm	174	87.1%	186	93%
Nervous	166	82.9%	180	90%
Angry	161	80.5%	157	78.5%
Total	501	83.5%	523	87.2%

5.3. Evaluation of E-alarm system

The E-alarm system will make a comprehensive emotional perception based on the real-time collected human posture images, and send e-mails corresponding to different emotions continuously. The posture collection module captures real-time human posture in 30 frames and displays the environmental images monitored by equipment on the screen. The emotion perception results are presented in emails with a photo of current environment. By using CNN for emotion perception, the working cycle of the system is 30s and meets the real-time requirements with high accuracy. The short working cycle (30s) enables users to make response to the emergencies in time.

The system is tested under three emotions for 120 times with 40 times of “calm”, 40 times of “nervous”, and 40 times of “angry”. The classification results and accuracy rates of the system under the three emotions are shown in **Table 2**. The number of correct emotion classifications are 110, and the accuracy is about 91.7%. The system shows a high accuracy of 97.5% when testing “calm” emotions. Only one wrong result fells on “nervous”. The accuracy of “nervous” and “angry” are 90% and 87.5%. Most of the misclassifications are caused by the confusion of these two emotions. The experiment results show that the E-alarm system has a high accuracy in the emotion classifications and can effectively perform emotional perception by collecting 30 frames of images and analyzing them.

Table. 2 Classification results of E-alarm system

Emotion	Test times	Correct results	Accuracy
Calm	40	39	97.5%
Nervous	40	36	90%
Angry	40	35	87.5%
Total	120	110	91.7%

6. CONCLUSION

This paper designs the E-alarm system in order to solve the problem that the monitoring system's lack of early alarming function for indoor violent conflict. The system can be applied to indoor monitoring, human-computer interaction and other aspects. It extracts the time and frequency domain features of human posture continuously and uses CNN to realize the perception of emotions effectively. The alarm emails containing corresponding emotions will be sent to the staffs so that they can respond in time. The E-alarm system includes three modules: posture collection module, emotion perception module, and real-time alarming module. The accuracy of emotion perception module by CNN is about 87.2%. The system has good real-time performance with the alarming cycle of 30s, and the accuracy of emotion perception reaches 91.7%.

However, although Kinect can solve the problem of human overlapping, it still has large error between the measured and the actual value. Data preprocessing method reduces the error to a certain extent instead of eliminating the error. Besides, the discrimination of human emotions between nervous and angry is fuzzy due to the self-made dataset. The classification accuracy of CNN and SVM classifiers fluctuate greatly. In the future, in order to improve the accuracy of the system, it is required for us to use standard dataset to test the performance of the classifiers, put forward perfect classification criteria, and optimize the deep learning module.

Acknowledgements

The authors thanks anonymous reviewers for their valuable suggestions and comments that have helped improve the presentation of this paper. This work was supported in part by the Thousand Talents Plan (the State Recruitment Program of Global Experts) under Grant No. WQ20141100198, in part by the Beijing Municipal Natural Science Foundation under Grant No. 3192028, and in part by the Beijing Municipal Natural Science Foundation under Grant No. L191020.

REFERENCES:

[1] V. Gentile, F. Milazzo, S. Sorce, A. Gentile, A. Augello and G. Pilato, "Body Gestures and Spoken Sentences: A Novel Approach for Revealing User's Emotions," in 2017 IEEE 11th International Conference on Semantic Computing (ICSC), 2017, pp. 69-72,2017.

[2] T. M. W. Vithanawasam and B. G. D. A. Madhusanka, "Dynamic Face and Upper-Body Emotion Recognition for Service Robots," 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS), 2018, pp. 428-432, 2018.

[3] S. Agrawal, R. Rangnekar, A. Das, S. Gawde and S. Dhage, "Gauging Customer Interest Using Skeletal Tracking and Convolutional Neural Network," 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2019, pp. 1-6, 2019.

[4] H. Zhang, T. Takahashi, Y. Kageyama, and M. Nishida, "Emotion Discrimination of Amusement Based on Three-Dimensional Data of Body Movements," International Journal of the Society of Materials Engineering for Resources, pp. 189-194, 2018.

[5] A. S. Patwardhan, "Multimodal mixed emotion detection," 2017 2nd International Conference on Communication and Electronics Systems (ICCES), 2017, pp. 139-143, 2017.

[6] Y. Maret, D. Oberson and M. Gavrilova, "Real-Time Embedded System for Gesture Recognition," 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2018, pp. 30-34, 2018.

[7] F. Ahmed, A. S. M. H. Bari and M. L. Gavrilova, "Emotion Recognition From Body Movement," in IEEE Access, vol. 8, pp. 11761-11781, 2020.

[8] M. A. Razzaq, J. Bang, S. S. Kang and S. Lee, "UnSkEm: Unobtrusive Skeletal-based Emotion Recognition for User Experience," 2020 International Conference on Information Networking (ICOIN), 2020, pp. 92-96, 2020.

[9] L. Wenyang, M. Xing, and Mu. J. Modern Electronic Technology, "Fall Behavior Detection and Analysis Based on Kinect V2," vol. 42, no. 06, pp. 150-153, 2019.

[10] B. Li, C. Zhu, S. Li, and T. J. I. T. o. A. C. Zhu, "Identifying Emotions from Non-contact Gaits Information Based on Microsoft Kinects," pp. 585-591, 2018.

[11] H. Li, J. Chen, and R. Hu, "Multiple feature fusion in convolutional neural networks for action recognition," Wuhan University Journal of Natural Sciences, pp. 73-78, 2017.

[12] X. Jianguo, X. Haifeng, and Z. Hua, "Text Classification Algorithm Based on Multi-Instance Learning Framework," Computer Engineering and Design, pp. 1017-1023, 2020.