Personnel Emotion Recognition Model for Internet of Vehicles Security Monitoring

Erkang Fu

Sichuan Agricultural University

Xi Li (✉ lixli@sicau.edu.cn)

Sichuan Agricultural University <https://orcid.org/0000-0002-3108-2091>

Zhi Yao

South China Normal University <https://orcid.org/0000-0002-4421-3804>

Yuxin Ren

Sichuan Agricultural University

Yuanhao Wu

Sichuan Agricultural University

Qiqi Fan

Sichuan Agricultural University

Research

Keywords: Emotion recognition, GSR, Convolution network.

Posted Date: August 17th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-797632/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Personnel emotion recognition model for Internet of vehicles security monitoring

Erkang Fu¹, Xi Li^{1*}, Zhi Yao², Yuxin Ren¹, Yuanhao Wu¹ and Qiqi Fan¹

¹ College of Landscape Architecture, Sichuan Agricultural University, Chengdu, China

² Guangdong Provincial Key Laboratory of Optical Information Materials and Technology & Institute of Electronic Paper Displays, South China Academy of Advanced Optoelectronics, South China Normal University, Guangzhou, China

lixli@sicau.edu.cn

Abstract. In recent years, the Internet of vehicles (IOV) with intelligent networked automobiles as the terminal node has gradually become the development trend of the automotive industry and the research hot spot in related fields. This is due to its characteristics of intelligence, networking, low-carbon and energy saving. Real time emotion recognition for drivers and pedestrians in the community can be utilized to prevent fatigue driving and malicious collision, keep safety verification and pedestrian safety detection. This paper mainly studies the face emotion recognition model that can be utilized for IOV. Considering the fluctuation of image acquisition perspective and image quality in the application scene of IOV, the natural scene video similar to vehicle environment and its galvanic skin response (GSR) are utilized to make the testing set of emotion recognition. Then an expression recognition model combining codec and Support Vector Machine (SVM) classifier is proposed. Finally, emotion recognition testing is completed on the basis of Algorithm 1. The matching accuracy between the emotion recognition model and GSR is 82.01%. In the process of model testing, 189 effective videos are involved and 155 are correctly identified.

Keywords: Emotion recognition, GSR, Convolution network.

1 Introduction

With the development and integration of information technology, computer technology and automobile manufacturing industry, IOV proposed to improve the level of automobile intelligent driving is known to the public. Intelligent networked automobiles make the IOV technology in automotive industry a hot spot. Then IOV technology takes the moving automobile as the object of information perception, and greatly improves the safety performance of the automobile by strengthening global optimization and control [1]. IOV is the specific implementation and application of traditional Internet of things technology in the automotive field. And it can greatly improve the intelligence and efficiency of traffic management by wireless communication technology and intelligent information processing technology. Therefore, IOV technology

can realize the intelligent monitoring and decision-making of vehicle information to realize the intelligent control of vehicles [2].

The face recognition technology under the background of artificial intelligence has been developed and applied rapidly in many fields because of its wide application range, strong operability and rich information. At present, the applications of face recognition mainly include face detection, identity recognition and emotion recognition. This paper focuses on face emotion recognition technology which can be applied to vehicle environment. Although the accuracy of facial emotion recognition in vehicle environment is disturbed by many factors such as angle fluctuation and transmission quality, its data has the characteristics of high feature discrimination and strong expression ability. Therefore, emotion recognition technology has high research value in the field of IOV real-time monitoring applied to fatigue driving, safety verification, malicious collision and pedestrian safety detection [3].

Emotion recognition is different from automobile manufacturing, which is the product of second industrial revolution with a long development process. However, it has become a hot research field with its excellent performance and application value [4]. The early concept of emotion recognition was pointed out in “Affective Computing” by Professor Picard of the Massachusetts Institute of Technology [5]. The emotion of human is often expressed by facial expressions, voices, gestures. Some scholars have conducted emotion recognition and analysis for these aspects [6-8]. American psychologist Mehrabian believes that facial expressions have the strongest ability to transmit information, and can achieve a recognition accuracy of 55% [9]. We believe that the voice and posture of face are affected by subjective psychological factors, resulting in insufficient representation ability. The common facial expression recognition is static image recognition, but the prediction of emotion requires dynamic facial expression because of its persistent. In addition, the development of physiology has made the recognition of human emotions by physiological data a hot field. In 2001, Picard et al. utilized multi-dimensional physiological signals to realize five levels of emotion recognition [10]. Subsequently, a large number of scholars began to analyze and research on physiological data and video emotion [11, 12]. In 2006, Savran et al. utilized the International Affective Picture System (IAPS) as a stimulus material to construct a data set “2005 emotional database” containing facial data and physiological data [13]. Koelstra et al. utilized pictures and music as stimulus materials to obtain expression videos and physiological data, and established the current popular emotion data set “DEAP” [14]. Later, Soleymani and others utilized the stimulation of network resources to construct “MAHNOB HCT” data set containing facial details, audio and physiological data [15]. It can be seen that the correlation of physiological data and emotional video to complete emotion recognition had become one of the mainstream directions in related research fields [16].

The emotion recognition model proposed in this paper was constructed by the rules shown in Algorithm 1 and expression recognition model. Therefore, the real labels of testing set were needed to verify the performance of the emotion recognition model. There were three common ways to set labels of video emotion. The first was directly defined by the known experimental conditions of the subjects. The second was defined by the emotional self-description of subjects after experiment. The third was

defined based on the physiological data of the subjects during the video shooting. We think the third method was more reliable than the previous method. Because the specified experimental conditions may not be able to stimulate the corresponding emotions for everyone and self-description was easily disturbed by psychological factors. The testing set was obtained by the video of middle school students under light exercise to make the research have application value.

Therefore, the video emotion recognition process was mainly divided into three processes, the definition of video emotional label, the training of video expression recognition models and the recognition of video emotions.

2 Experiment and proposed method

2.1 Preparation of physiological data

At present, there are many video data sets about emotion recognition, including the early Cohn Kanade dataset plus (CK+) [17] and recent DEAP data set. These data sets have the advantages of rich content and strong representation. However, the above video data sets only have standard faces, which is very different from the facial video data sets with multi angles and large quality fluctuation under the background of IOV applications. Therefore, the facial video of young people's daily behavior was collected to study the face emotion recognition in the vehicle environment. In addition, the physiological data were obtained at the same time to complete accurate emotion prediction and inference. Fig. 1 shows the physiological data of the subjects in the video state.

	A	B	C	D	E	F	G
1	min	CH1	CH2	CH3	CH4	CH5	CH6
2		397000	397000	397000	397000	397000	397000
3	0.00000000	0.644531	3.97949	3.44238	-0.234375	-0.00244141	-0.00732422
4	0.00000833	0.644531	3.97949	3.44238	-0.234375	-0.00244141	0.00488281
5	0.00001667	0.654297	3.97949	3.44238	-0.239258	0.00244141	0.00976563
6	0.00002500	0.654297	3.97949	3.44238	-0.239258	0.00244141	0.00488281
7	0.00003333	0.668945	3.97949	3.44727	-0.241699	0.00488281	-0.00732422
8	0.00004167	0.668945	3.97949	3.44727	-0.241699	0	-0.00244141
9	0.00005000	0.683594	3.97949	3.44727	-0.246582	0.00244141	0.00976563
10	0.00005833	0.683594	3.97949	3.44727	-0.246582	0	0.0146484
11	0.00006667	0.693359	3.97949	3.45215	-0.253906	-0.00488281	-0.00244141
12	0.00007500	0.693359	3.97949	3.45215	-0.253906	-0.00732422	-0.00244141
13	0.00008333	0.708008	3.97949	3.45703	-0.258789	-0.00244141	-0.00244141
14	0.00009167	0.708008	3.97949	3.45703	-0.258789	0	0
15	0.00010000	0.717773	3.97949	3.45703	-0.266113	0	-0.00732422
16	0.00010833	0.717773	3.97949	3.45703	-0.266113	0.00488281	-0.00732422
17	0.00011667	0.732422	3.97949	3.46191	-0.275879	0.00732422	-0.00488281
18	0.00012500	0.732422	3.97949	3.46191	-0.275879	0.00488281	0
19	0.00013333	0.742188	3.97949	3.46191	-0.285645	0.00244141	-0.00732422
20	0.00014167	0.742188	3.97949	3.46191	-0.285645	0.00488281	-0.00732422
21	0.00015000	0.756836	3.97949	3.4668	-0.292969	0.00244141	-0.00488281
22	0.00015833	0.756836	3.97949	3.4668	-0.292969	0.00488281	0
23	0.00016667	0.766602	3.97949	3.4668	-0.300293	0	-0.00244141
24	0.00017500	0.766602	3.97949	3.4668	-0.300293	0	-0.00732422
25	0.00018333	0.78125	3.97949	3.47168	-0.305176	0	0.00244141
26	0.00019167	0.78125	3.97949	3.47168	-0.305176	-0.00244141	0.00488281
27	0.00020000	0.791016	3.97949	3.47168	-0.307617	0	-0.00244141

Fig. 1. Experimental data set-partial physiological data

The first channel is Heart rate based on PhotoPlethysmoGraphy (PPG). The second channel is the value of GSR. The third channel is the value of electrical signal of respiration. The fourth channel is the value of ElectroCardioGram (ECG). The fifth channel is the value of ElectroEncephaloGram (EEG).

2.2 The definition of video emotional label

The GSR of human is controlled by the human nervous system, It has a strong ability to characterize physiological characteristics [18]. A large number of studies have shown that emotional fluctuations can cause significant changes in GSR [19, 20]. Therefore, GSR is selected to define video emotional label to improve work efficiency. Related studies utilizes the feature extraction method of the University of Augsburg in Germany to find that the emotion of subject can reflect the characteristics on GSR. This conclusion can be shown in Fig. 2 [21].

Fig. 2. The relationship between common emotions and GSR

Above images are derived from the characteristic results of some subjects and they are not very persuasive. However, the following applicable conclusion can be obtained by observation and testing of data set when emotion levels are categorized into three categories.

- Happy: Within the range of video, there are denser multi-band peaks, which are mostly distributed at the beginning of the video;
- Quiet: Within the range of video, there is basically no crest or only once at both ends;
- Unhappy: Within the range of video, there are crests at the beginning and end of the video, or only dense crests appear in the middle of the video with almost no intervals.

After the above rules are summarized, the emotional label of the testing video is defined by the value of GSR and verification. The specific experimental steps are as follows.

Data preprocessing. Firstly, the most representative GSR in the physiological data is completed noise reduction and smoothness. The abnormal value in the data is changed to a nearby value to complete data noise reduction. Savitzky-Golay filter is utilized to smooth the data. The Savitzky-Golay filter is a digital filter that fits adjacent data points to a low-order polynomial by the linear least squares [22]. The solution of least squares equation can be found when the data spacing is equal. Fig. 3 is a diagram of its smoothing process.

Fig. 3. The principle of Savitzky-Golay filtering

The blue point in each window in Fig. 3 is the center point of the window, and the mathematical principle of filtering is (1).

$$x_{k,smooth} = \bar{x} = \frac{1}{H} \sum_{i=-w}^{+w} x_{k+i} h_i \quad (1)$$

The Savitzky-Golay filter utilizes the least squares to regress a small window of data to a polynomial, and then utilizes the polynomial to estimate the point at the center of window. Where h_i is the smoothing coefficient. $\frac{h_i}{H}$ is fitted by the principle of least squares in (1).

On the same curve, different window widths can be selected at any position to meet the needs of different filtering. This is useful for processing time series data at different stages.

The definition of emoticons in videos. Emotional swings are short and continuous when they are not stimulated by a strong external environment. Therefore, the definition method of dividing time segment is utilized to define the emotional label of each short video. The total length of each video is about 3min. It includes the process from the beginning of experiment to the completion of the specified action and then the end

of experiment. Therefore, we believe that this process can reflect a variety of specific emotions. It is stipulated that every 15s video is defined as an emotional video to improve the accuracy of definition. There is a 5s interval between every two emotional videos. Then the emotional label of each short video is defined based on the relationship from Fig. 2 and the value of GSR after preprocessing.

2.3 The definition of video emotional label

The expression recognition model of the paper is a combination of convolutional codec and SVM classifier. And the model and Algorithm 1 cooperate to complete the prediction of emotions. The model features extracted by the convolutional codec have strong abstraction. The feature can reduce the training noise caused by the large difference of facial style. The core of the codec is image convolution and image deconvolution.

Image convolution is developed from signal convolution. Image convolution is obtained by expanding the one-dimensional signal in two dimensions and rotating its convolution kernel by 180°. Image convolution introduces the three calculation concepts of convolution kernel F , stride S , and padding P . Their calculation relationship is shown in (2) and (3) [23].

$$W' = (W - F + 2P)/S + 1, \quad (2)$$

$$H' = (H - F + 2P)/S + 1. \quad (3)$$

The above equations represent the calculation of the output image when the size of input image is $[W \times H \times D]$. Where W' and H' represent the width and height of output image, and the depth D' of the output image is determined by the number of convolution kernels.

In a convolutional network, reasonable settings of F , S , P are required to ensure that the size of image is controllable and the number of network layers continues to rise. Fig. 4 shows a common image convolution process.

Fig. 4. Schematic diagram of image convolution

The value of S is 1 in the convolution process shown in Fig. 4. A total of four convolutions occur in the convolution shown in Fig. 4. The process of convolution can be

digitized when the image is expanded as shown in Fig. 5 and the convolution kernel is expanded into matrix.

Fig. 5. Matrix representation of the image

The matrix of the convolution kernel is shown in (4).

$$\begin{bmatrix} w_{0,0} & w_{0,1} & w_{0,2} & 0 & w_{1,0} & w_{1,1} & w_{1,2} & 0 & w_{2,0} & w_{2,1} & w_{2,2} & 0 & 0 & 0 & 0 & 0 \\ 0 & w_{0,0} & w_{0,1} & w_{0,2} & 0 & w_{1,0} & w_{1,1} & w_{1,2} & 0 & w_{2,0} & w_{2,1} & w_{2,2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & w_{0,0} & w_{0,1} & w_{0,2} & 0 & w_{1,0} & w_{1,1} & w_{1,2} & 0 & w_{2,0} & w_{2,1} & w_{2,2} & 0 \\ 0 & 0 & 0 & 0 & 0 & w_{0,0} & w_{0,1} & w_{0,2} & 0 & w_{1,0} & w_{1,1} & w_{1,2} & 0 & w_{2,0} & w_{2,1} & w_{2,2} \end{bmatrix}. \quad (4)$$

Therefore, the operation of convolution can be expressed by (5).

$$Y = CX. \quad (5)$$

Where Y represents the result of convolution, and C represents the matrix of the convolution kernel. First convolution is the multiplication of the first line of (4) with the matrix of image in (5), and subsequent convolution is also based on this process of calculation. A vector with size of $[4 \times 1]$ is subsequently obtained after the calculation of (5). The output image after convolution can be restored by following the reverse process of Fig. 5. Therefore, the convolution process can be described as a multiplication of weight matrix with an image vector.

The process of convolution is essentially a combination of forward propagation and backward derivative propagation. The principle of deriving x in back propagation is (6).

$$\frac{\partial Loss}{\partial x_i} = \sum_i \frac{\partial Loss}{\partial y_i} \times \frac{\partial y_i}{\partial x_i}. \quad (6)$$

Where y_i can be expressed by (7).

$$y_i = \sum_{j=1}^{16} C_{ij} X_j. \quad (7)$$

Then (8) can be obtained by (7).

$$\frac{\partial y_i}{\partial x_j} = C_{ij}. \quad (8)$$

And (9) can be obtained by substituting (8) into (6).

$$\frac{\partial Loss}{\partial x_i} = \sum_{i=1}^4 \frac{\partial Loss}{\partial y_i} \times C_{ij}. \quad (9)$$

The multiplication of matrices can be achieved by changing Σ in (9) to the form of a matrix.

$$\frac{\partial Loss}{\partial x_j} = \left(\frac{\partial Loss}{\partial y} \right)^T \times C_{*j} = C_{*j}^T \times \left(\frac{\partial Loss}{\partial y} \right). \quad (10)$$

Where C_{ij} is the matrix of forward propagation, and C_{*j} is the matrix of backward propagation. In the process of deconvolution, the mathematical meanings of the above two parameters need to be exchanged. The relationship of calculation corresponding to deconvolution is shown in (11) and (12), which means that W and H before convolution are obtained by inverse calculation of W' and H' after convolution.

$$W = S(W' - 1) - 2P + F, \quad (11)$$

$$H = S(H' - 1) - 2P + F. \quad (12)$$

Therefore, the structure of the expression recognition model is shown in Fig. 6.

Fig. 6. The structure of the expression recognition model

Where Conv represents for convolution and Deconv represents for deconvolution.

2.4 The training and testing of the expression recognition model

The initialization state of the testing set is unlabeled to ensure the rationality of emotional label. Therefore, another data set with self-descriptive labels is utilized to com-

plete the training of the model. Fig. 7 shows partial data set utilized to train the expression recognition model.

Fig. 7. The training set of the expression recognition model

Therefore, parameter values of the model can be obtained by training the model on the data set shown in Fig. 7. Subsequent testing follows the principle of video frame image analysis. Nearly 30 frames images appear each 1s in the testing video. Following recognition process is defined so that the label of the expressions can be accurately defined within 1s.

CascadeClassifier in Opencv is utilized for face detection in the process of model testing. This is a cascaded classifier utilizing Harr feature of images. The principle of Harr feature can be utilized to complete face recognition well [24]. Fig. 8 shows the result of face detection and the expression recognition model.

Fig. 8. Partial result of face detection and the expression recognition model

Therefore, the recognition result of the expression in each frame image can be obtained. Fig. 8 shows that a face is detected by CascadeClassifier and it is recognized as happy by the expression recognition model.

The most important thing in this section is that the result of expression recognition of each frame image is transformed into the emotion recognition result of the image per second. The emotion in a second is considered to be 'Unhappy' if the number of 'Unhappy' in 1 this second is greater than 6 and the number of 'Happy'. If not, the judgments of other emotions are subsequently continued. The emotion in a second is considered to be 'Happy' if the number of 'Happy' in 1 second is greater than 4 and the number of 'Unhappy'. And the emotion in a second is considered to be 'Happy' if the number of 'Happy' in 1 second is greater than 0 and the number of 'Unhappy' when the number of 'Quiet' in 1 second is less than 5. The emotion per second is defined as calm when the above conditions are not met.

2.5 Emotion recognition for short videos

Emotion sequence represented by '1', '0' and '2' can be obtained after the expression prediction model is trained and tested. The above three numbers represent three emotions defined in Section 2.2. The emotion prediction of each short video is completed by Algorithm 1 after the emotional label per second is obtained.

Algorithm 1 Rule for definition of emotion

input: Total number N , Number of '0' N_0 , Number of '1' N_1 , Number of '2' N_2 , Number of '0' per 15s n_0 , Number of '1' per 15s n_1 , Number of '2' per 15s n_2 , Number of '1' per 5s n_{i1} , Number of '2' per 5s n_{i2} , Distance d_{11} , Distance d_{12}

output: the emotion of the subjects per 15s E

- 1 **if** $N_1 < 3$ or $N_2 < 3$ **then**
 - 2 The subjects have no emotional changes in this test
 - 3 **if** $n_1 = 0$ and $n_2 > 1$ **then**
 - 4 $E = \text{unhappy}$
 - 5 **if** $n_1 > 5$ or '1' occurs four times in a row per 15s **then**
 - 6 $E = \text{happy}$
 - 7 **if** $N_1 < 70$ and $n_{i1} > 2$ and $d_{11} < d_{12}$ **then**
 - 8 $n_1 = n_1 + \text{former } n_{i1}$
 - 9 **if** $n_1 > 7$ **then**
 - 10 $E = \text{happy}$
 - 11 **else if** $N_1 < 70$ and $n_{i1} > 2$ and $d_{11} > d_{12}$ **then**
 - 12 $n_1 = n_1 + \text{latter } n_{i1}$
 - 13 **if** $n_1 > 7$ **then**
 - 14 $E = \text{happy}$
 - 15 **if** $N_1 < N$ and $N_1 > \frac{1}{5}N$ and $n_1 > 4$ **then**
 - 16 $E = \text{happy}$
 - 17 **else if** $n_{i1} > 2$ and $d_{11} < d_{12}$ **then**
 - 18 $n_1 = n_1 + \text{former } n_{i1}$
 - 19 **if** $n_1 > 7$ **then**
-

```

20   E= happy
21   else if  $n_{i1} > 2$  and  $d_{11} > d_{12}$  then
22      $n_1 = n_1 + \text{latter } n_{i1}$ 
23     if  $n_1 > 7$  then
24       E= happy
25   if  $N_1 < \frac{1}{5}N$  and  $N_1 > \frac{1}{6}N$  and  $n_1 > 3$  then
26     E= happy
27   if  $N_1 < \frac{1}{6}N$  and  $N_1 > 12$  and  $n_1 > 2$  then
28     E= happy
29   if  $N_2 < \frac{1}{6}N$  and  $N_2 > 4$  and  $n_2 > 1$  then
30     E= unhappy
31   if  $n_1 > 0$  and  $n_2 > 0$  and  $n_1 > n_2$  then
32      $n_1 = n_1 + n_2$ 

```

Where N represents the total number of emoticons per second in the video. N_0, N_1, N_2 respectively indicate the number of corresponding emotions. n_0, n_1, n_2 are utilized to represent the number of different emoticons in the video each 15 seconds. The number of 'Happy' and 'Unhappy' is defined as n_{i1} and n_{i2} of interval video in each 5 second to reduce the loss of useful information. In addition, the distance between first '1' appearing in each interval video and last '1' in 15s video of previous section is defined as d_{11} . The distance between last '1' appearing in each interval video and first '1' in 15s video of previous section is defined as d_{12} . Therefore, the emotion recognition model needs the characteristics of different subject analyzed by Algorithm 1.

2.6 The overall structure of the emotion recognition model

The overall flow chart of the model is shown in Fig. 9 to clearly show the process of emotion recognition.

Fig. 9. The overall structure of the emotion recognition model

3 Results and discussion

3.1 The result of preparation

The data set utilized for training needed to be grayed and standardized. The preparation of data set utilized for testing was divided into two parts. The part of video needed to be grayed and standardized, and the part of GSR needed to be completed data noise reduction and smoothness. Partial result of the latter was shown in Fig. 10 and Fig. 11.

Fig. 10. The value of GSR during experiment for a subject

Fig. 11. The value of GSR after preprocessing for a subject

3.2 The setting of parameter

The training parameters of the proposed expression recognition model and comparison models were shown in Table 1.

Table 1. Training parameters.

Parameter	Value
batch_size	32
Image_size	112×112
learning_rate	1e-3

This parameter model was mainly applied to the proposed facial expression recognition model and its comparison model: Resnet18 and VGG16. The initial weights obtained by transfer learning on ImageNet were applied to its comparison model.

The parameters of CascadeClassifier were shown in Table 2.

Table 2. Parameters of CascadeClassifier.

Parameter	Value
expansion scale	160×160
Scale_factor	7
Min_size	100×100
Max_size	1000×1000

3.3 Model evaluation method

The evaluation of the model in the paper was obtained by matching the label defined by GSR with the result of emotion recognition model. The former was utilized as the true value of video emotion, the latter was the predicted value. (13) for model evaluation was as follows.

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (13)$$

Where TP referred to predict the target class as the number of target classes, TN referred to predict the non-target class as the number of non-target classes, FN referred to predict the non-target class as the number of target classes, FP referred to predict the target class as the number of non-target classes.

3.4 Experimental results.

The results of the emotion recognition model were mainly divided into two parts: the training part and testing part. Fig. 12 showed the training part of the model.

Fig. 12. The training process of the proposed model

The loss decay of the model and the increase of accuracy took a total of 120 epochs. The expressive ability of feature extracted by the model was very strong alby the training process of the model was slow. The original image can be restored more accurately by this feature. The input images were uniformly grayed and standardized to reduce the influence of facial background. The comparison between original image and generated image from its features was shown in Fig. 13.

Fig. 13. Original image and generated image from its features

SVM classifier needed to be retrained to complete the intact process of model training after the abstract features were obtained by the codec model. The training of SVM classifier included the selection of its kernel function γ and penalty coefficient c . The selection of the kernel function was shown in Fig. 14.

Fig. 14. The classification accuracy of SVM on testing set with different kernel and value of γ

As shown in Fig. 14, it can be found that the highest accuracy was 97.71% when the kernel function is gauss. It can also be obtained that the peak accuracy was about 97.45% and 91.59%, respectively when the kernel function is poly and sigmoid. And it can be determined that best values of γ corresponding to three kernel functions were 8.935, 8.935, 2, respectively. On this basis, the process of adjusting the penalty coefficient c was shown in Fig. 15.

Fig. 15. The accuracy of classification changed with the value of c

It can be found that the highest accuracy of classification was 97.83% when kernel function was gauss, the value of γ was 8.935 and the value of c was 15.

Resnet18 and VGG16 were trained and verified on the training set to further illustrate the superiority of the proposed model in expression recognition. Fig. 16 showed the result of training and verification of the comparison model on training set.

Fig. 16. The training process of Resnet18 and VGG16

It can be found from Fig. 16 that the accuracy of Resnet18 was slightly better than VGG16, and its highest accuracy of verification reached 97.64% when its epoch was 11. But this result was also slightly lower than the proposed model.

The predicted results of the models were matched with the labels defined by GSR after the training of three models. The matching accuracy was 82.01% after testing all subjects. The number of effective short video emoticons in testing set was 189, and the number correctly identified was 155. The testing accuracy of the comparison models were both 69.84%. The testing accuracy of three models changed with the increase in the number of subjects was shown in Fig. 17.

Fig. 17. The diagram of testing accuracy

The three curves quickly reached peaks and their accuracy decreased in the second half of them. A confusion matrix was utilized to further show the result of the proposed emotion recognition model to analyze the limitation of model. The matrix was shown in Fig. 18.

Fig. 18. Confusion matrix of the proposed emotion recognition model

Following conclusions can be obtained combined with the detailed result of emotion recognition in Fig. 17 and Fig. 18.

- The differences of facial styles led to the difficulty of facial emotion recognition in some subjects with unclear facial expressions or rich facial expressions. This class of subjects in testing set distributed in the second half of the data set. This conclusion was also reflected in the misjudgment in Fig. 18;
- The model can not detect facial expression due to the subjects' head down, which led to the decrease of accuracy;
- The recognition of 'Quiet' emotion in natural scene had high recognition accuracy due to the high frequency of 'Quiet' emotion. However, the ability of recognition for 'Unhappy' emotion was weak due to the small number of samples.

4 Conclusions

Firstly, The reliable emotional label of the proposed emotion recognition model was obtained from GSR. Then the model achieved an accuracy of 82.01% by a reasonable process of recognition. The proposed model has certain practical value for predicting the human emotion of natural activities in vehicle environment due to its data set utilized in the model has certain characteristic similarity with the video data in the vehicle environment, .

List of abbreviations

IOV: Internet of vehicles; GSR: galvanic skin response; SVM: Support Vector Machine; IAPS: International Affective Picture System; CK+: Cohn Kanade dataset plus; PPG: PhotoPlethysmoGraphy; ECG: ElectroCardioGram; EEG: ElectroEncephaloGram

Declarations

Availability of data and materials

The datasets used and analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

Supported by project of shenzhen science and technology innovation committee (JCYJ20190809145407809), project of shenzhen Institute of Information Technology School-level Innovative Scientific Research Team (TD2020E001).

Authors' contributions

Erkang Fu and Xi Li designed the research. Zhi Yao and Yuxin Ren conducted the literature review and wrote this manuscript. Yuanhao Wu and Qiqi Fan performed the numerical calculations and derived the formulae in the paper. All authors contributed to the literature review, discussion of the results and edited the manuscript.

Acknowledgements

I would like to thank Wang Yihao, Feng Jinyu, Cen Muosuo, Liu Renwei and Wang Min for their hard work in the data collection stage.

Authors' information

¹College of Landscape Architecture, Sichuan Agricultural University, 611130, Chengdu, China.

²Guangdong Provincial Key Laboratory of Optical Information Materials and Technology & Institute of Electronic Paper Displays, South China Academy of Advanced Optoelectronics, South China Normal University, 510006, Guangzhou, China.

References

1. SS Devi, A Bhuvaneswari, Quantile Regressive Fish Swarm Optimized Deep Convolutional Neural Learning for Reliable Data Transmission in IoV. *International journal of Computer Networks and Communications*. 13(2), 81-97 (2021). doi:10.5121/ijcnc.2021.13205
2. F Valocky, M Orgon, I Fajdiak, Experimental Autonomous Car Model with safety sensor in Wireless Network. *IFAC PapersOnLine*. 52(27), 92-97 (2019). doi:10.1016/j.ifacol.2019.12.739
3. K Afzal, R Tariq, F Aadil, Z Iqbal, M Sajid, An Optimized and Efficient Routing Protocol Application for IoV. *Mathematical Problems in Engineering*. (2021). doi:10.1155/2021/9977252
4. S Turabzadeh, HY Meng, RM Swash, M Pleva, J Juhar, Facial Expression Emotion Detection for Real-Time Embedded System. *Technologies*. 6:17 (2018). doi:10.3390/technologies6010017
5. RW Picard, *Affective computing: challenges*. (USA: MIT Press, 1997), pp. 2-10

6. K Anderson, PW Mcowan, A real-time automated system for the recognition of human facial expressions. *IEEE TRANSACTIONS ON CYBERNETICS*. 36, 96-105 (2006). doi:10.1109/TSMCB.2005.854502
7. J Ang, R Dhillon, A Krupski, E Shriberg, A Stolcke, Prosody-based automatic detection of annoyance and frustration in human-computer dialog. *Seventh International Conference on Spoken Language Processing*. 2037-2040 (2002).
8. C Feichtenhofer, A Pinz, A Zisserman, Convolutional Two-Stream Network Fusion for Video Action Recognition. *Computer Vision and Pattern Recognition*. IEEE. 1933-1941 (2016). doi:10.1109/CVPR.2016.213
9. A Mehrabian, Communication without words. *Psychology Today*. 2, 53-55 (1968). doi:10.1016/S0140-6736(65)90194-7
10. RW Picard, E Vyzas, J Healey, Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 23, 1175-1191 (2001). doi:10.1109/34.954607
11. X Liu, XY Zhang, NOMA-based Resource Allocation for Cluster-based Cognitive Industrial Internet of Things. *IEEE Transactions on Industrial Informatics*. 16, 5379-5388 (2020). doi:10.1109/TII.2019.2947435
12. N Samadiani, G Huang, W Luo, CH Chi, YF Shu, R Wang, T Kocaturk, A multiple feature fusion framework for video emotion recognition in the wild. *Concurrency and Computation: Practice and Experience*. (2020). doi:10.1002/cpe.5764
13. A Savran, K Ciftci, G Chanel, J Mota, L Viet, B Sankur, L Akarun, A Caplier, M Rombaut, Emotion Detection in the Loop from Brain Signals and Facial Images. *International Summer Workshop on Multimodal Interfaces*. (2006). doi:10.17660/ActaHortic.2005.671.18
14. S Koelstra, C Muhl, M Soleymani, JS Lee, Yazdani A, Ebrahimi T, Pun T, Nijholt A, Patras I. Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*. 3, 18-31 (2012). doi:10.1109/T-AFFC.2011.15
15. M Soleymani, J Lichtenauer, T Pun, M Pantic, A Multimodal Database for Affect Recognition and Implicit Tagging. *IEEE Transactions on Affective Computing*. 3, 42-55 (2012). doi:10.1109/T-AFFC.2011.25
16. WR Hu, G Huang, LL Li, L Zhang, ZG Zhang, Z Liang, Video-triggered EEG-emotion public databases and current methods: A survey. *Brain Science Advances*. 6, 255-287 (2019). doi:10.26599/BSA.2020.9050026
17. P Lucey, JF Cohn, T Kanade, J Saragih, Z Ambadar, I Matthews, The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. *Proceedings of the Third International Workshop on CVPR for Human Communicative Behavior Analysis(CVPR4HB)*. 94-101 (2010). doi:10.1109/CVPRW.2010.5543262
18. C Tronstad, H Kalvøy, S Grimnes, G Martinsen Ørjan, Improved Estimation of Sweating Based on Electrical Properties of Skin. *Annals of biomedical engineering*. 41, 1074-1083 (2013). doi:10.1007/s10439-013-0743-4
19. MM Bradley, PJ Lang, Measuring emotion: Behavior, feeling, and physiology. Lane RD, Nadel L, editors. *Cognitive neuroscience of emotion*. (New York: Oxford University Press, 2000), pp. 242-276
20. PJ Lang, Emotion and Motivation: Attention, Perception, and Action. *Journal of Sport & Exercise Psychology*. 22, 180-199 (2020). doi:10.1097/00005131-200006000-00017
21. KH Kim, SW Bang, SR Kim, Emotion recognition system using short-term monitoring of physiological signals. *Medical & biological engineering & computing*. 42, 419-427 (2004). doi:10.1007/BF02344719

22. HH Madden, Comments on Savitzky-Golay convolution method for least-squares fit smoothing and differentiation of digital data. *Analytical Chemistry - ANAL CHEM.* 50, 1383-1386 (1978). doi:10.1021/ac50031a048
23. V Dumoulin, F Visin, A guide to convolution arithmetic for deep learning. (2019), pp. 1-28
24. R Lienhart, A Kuranov, V Pisarevsky, Empirical analysis of detection cascades of boosted classifiers for rapid object detection. *Joint Pattern Recognition Symposium.* 2781, 297-304 (2003). doi:10.1007/978-3-540-45243-0_39