EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

Personnel emotion recognition model for Internet of vehicles security monitoring in community public space

Erkang Fu[1], Xi Li[1*], Zhi Yao[2], Yuxin Ren[1], Yuanhao Wu[1] and Qiqi Fan[1]

*Correspondence: lixi@sicau.edu.cn
[1] College of Landscape Architecture, Sichuan Agricultural University, Chengdu 611130, China
Full list of author information is available at the end of the article

Abstract

In recent years, the Internet of vehicles (IOV) with intelligent networked automobiles as terminal node has gradually become the development trend of automotive industry and 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 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, the IOV proposed to improve the level of automobile intelligent driving is known to public. IOV is a branch of industrial Internet of things (IOT) technology, so it also has the advantages of sensing technology, mobile communication technology and intelligent analysis of the IOT [1]. 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 [2]. IOV is the specific implementation and application of traditional Internet of things technology in 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 [3].

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. Community is an important part of a city, but due to the lack of intelligent means in the traditional community management mode, it can not meet the residents' needs for safe and efficient community service. This paper focuses on face emotion recognition technology which can be applied to vehicle environment in community public space. 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 [4].

Emotion recognition is different from automobile manufacturing, the latter 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 [5]. The early concept of emotion recognition was pointed out in "Affective Computing" by Professor Picard of the Massachusetts Institute of Technology [6]. The emotion of human was often expressed by facial expressions, voices, gestures. Some scholars had conducted emotion recognition and analysis for these aspects [7–9]. American psychologist Mehrabian believed that facial expressions have the strongest ability to transmit information, and they can be utilized to achieve a recognition accuracy of 55% in emotion recognition [10]. We believed that the voice and posture of face are affected by subjective psychological factors, which leads to insufficient representation ability. The common facial expression recognition was static image recognition, but the prediction of emotion required dynamic facial expression because of its persistence. In addition, the development of physiology had 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 [11]. Subsequently, a large number of scholars began to analyze and research on physiological data and video emotion [12, 13]. 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 [14]. Koelstra et al. utilized pictures and music as stimulus materials to obtain expression videos and physiological data, then they established the current popular emotion data set "DEAP" [15]. Later, Soleymani and others utilized the stimulation of network resources to construct "MAHNOB HCI" data set containing facial details, audio and physiological data [16]. It can be seen that research on the correlation of physiological data and video emotion to complete emotion recognition had become one of the mainstream directions in related fields [17]. In addition, a large number of physiological and emotional data brought high load, high power consumption and resource shortage to the IOV system. And Fifth generation (5G) network can be well applied to the communication and transmission of IOT with its sufficient spectrum resources. Therefore, a large number of scholars had studied and analyzed the optimization of 5G communication technology and its combination with IOT 、IOV [18, 19].

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 labels of video emotion in first method was directly defined by the known experimental conditions of subjects. The labels of video emotion in second method was defined by the emotional self-description of subjects after experiment. The label of video emotion in third method was defined based on the physiological data of the subjects during video shooting. We thought the third method was more reliable than the previous method. Because 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 subjects under the natural scene video similar to vehicle environment to make the research more valuable.

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 emotion.

2 Experiment and proposed method

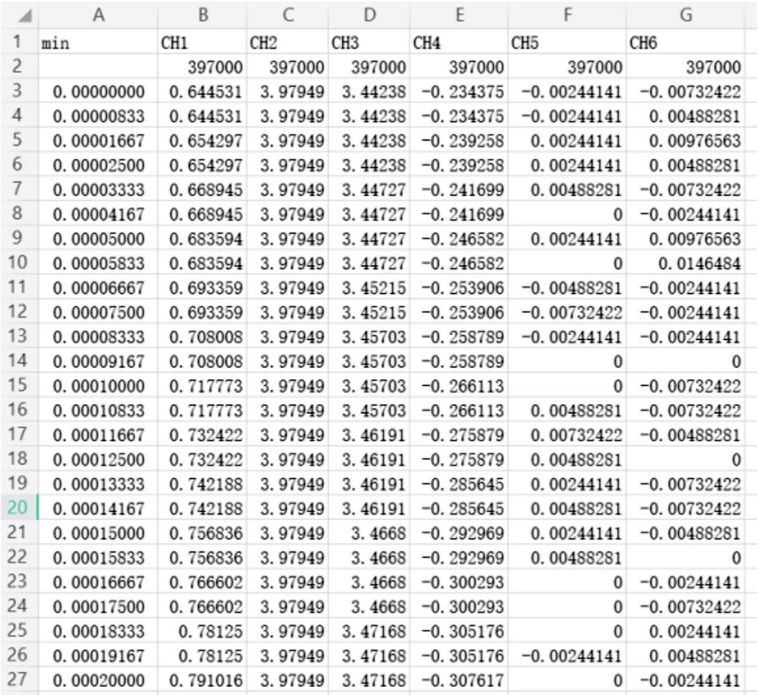
2.1 Preparation of physiological data

At present, there are many video data sets about emotion recognition, including the early Cohn Kanade dataset plus (CK+) [20] 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 specific behavior is collected to study the facial emotion recognition in the vehicle environment. In addition, the physiological data are obtained to complete accurate emotion prediction and inference. Figure 1 shows the physiological data of the subjects in the video state.

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 human nervous system, it has strong physiological characteristics [21]. A large number of studies have shown that emotional fluctuations can cause significant changes in GSR [22, 23]. Therefore, GSR is selected to define video



	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 Partial physiological data of experimental data set

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 subjects can reflect their characteristics on GSR. This conclusion can also be shown in Fig. 2 [24].

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 levels of emotion 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 peaks or only once at both ends;
- Unhappy: Within the range of video, there are peaks at the beginning and end of the video, or only dense peaks appear in the middle of the video with almost no intervals.

After 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 smoothing. The abnormal value in the data is updated to its 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

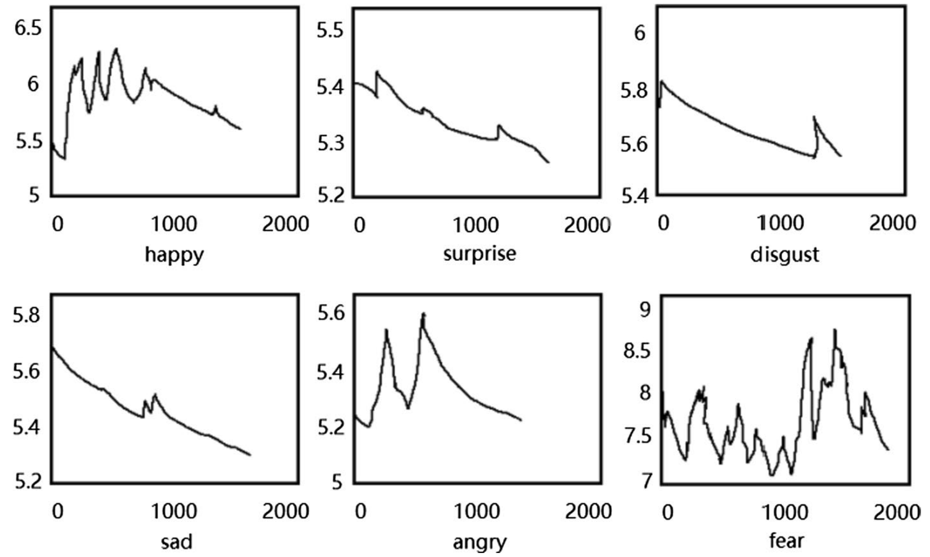


Fig. 2 The relationship between common emotions and GSR

to a low-order polynomial by linear least squares [25]. The solution of least squares equation can be found when the data spacing is equal. Figure 3 is a diagram of its smoothing process.

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

$$x_{k,\text{smooth}} = \overline{x} = \frac{1}{H}\sum_{i=-w}^{+w} x_{k+i}h_i. \tag{1}$$

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 widths of window 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 emotions 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 3 min. 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 15 s video is defined as an emotional video to improve the accuracy of definition. There is a 5 s interval between every two emotional videos.

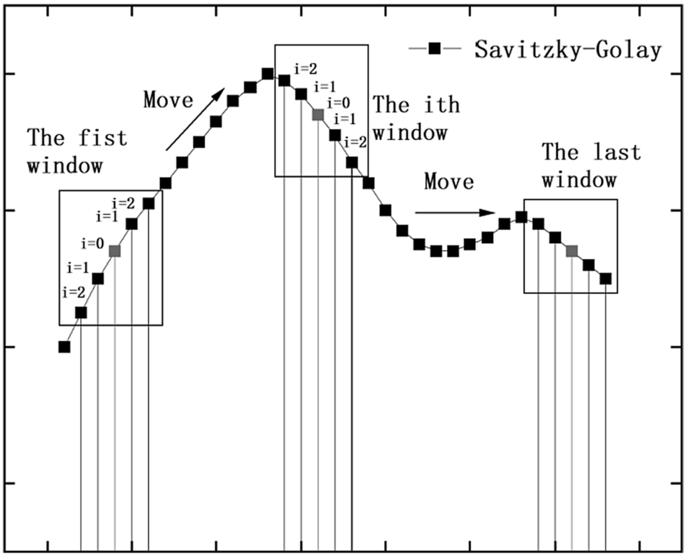


Fig. 3 The principle of Savitzky–Golay filtering

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 Structure and principle of the proposed model

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 video emotion. 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) [26].

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

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

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.

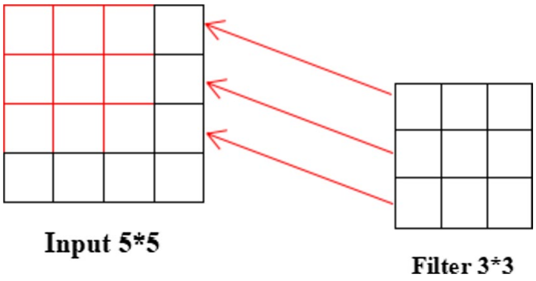


Fig. 4 Schematic diagram of image convolution

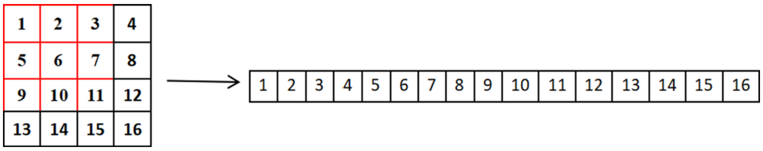


Fig. 5 Matrix representation of the image

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. Figure 4 shows a common image convolution process.

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.

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

$$\begin{bmatrix} w_{0,0} & w_{0,1} & w_{0,2} & 0 & w_{1,0} & w_{1,1} & w_{1,2} & 0 & w_{2,0} & w_{2,1} & w_{2,2} & 0 & 0 & 0 & 0 & 0 \\ 0 & w_{0,0} & w_{0,1} & w_{0,2} & 0 & w_{1,0} & w_{1,1} & w_{1,2} & 0 & w_{2,0} & w_{2,1} & w_{2,2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & w_{0,0} & w_{0,1} & w_{0,2} & 0 & w_{1,0} & w_{1,1} & w_{1,2} & 0 & w_{2,0} & w_{2,1} & w_{2,2} & 0 \\ 0 & 0 & 0 & 0 & 0 & w_{0,0} & w_{0,1} & w_{0,2} & 0 & w_{1,0} & w_{1,1} & w_{1,2} & 0 & w_{2,0} & w_{2,1} & w_{2,2} \end{bmatrix}. \tag{4}$$

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

$$Y = CX. \tag{5}$$

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.

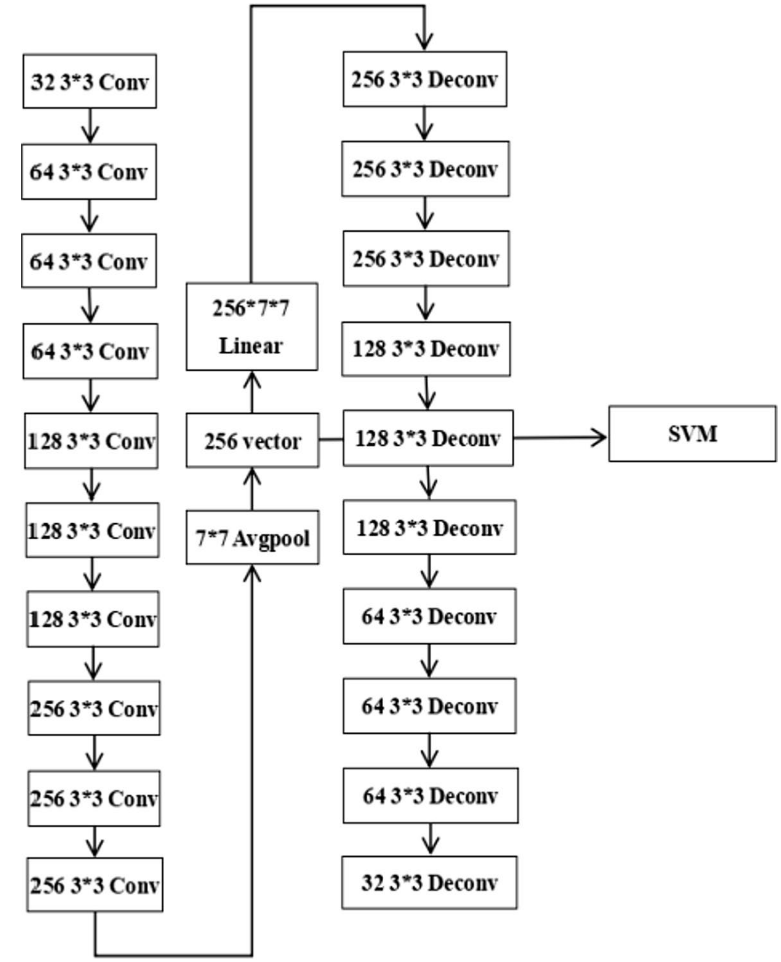


Fig. 6 The structure of the expression recognition model

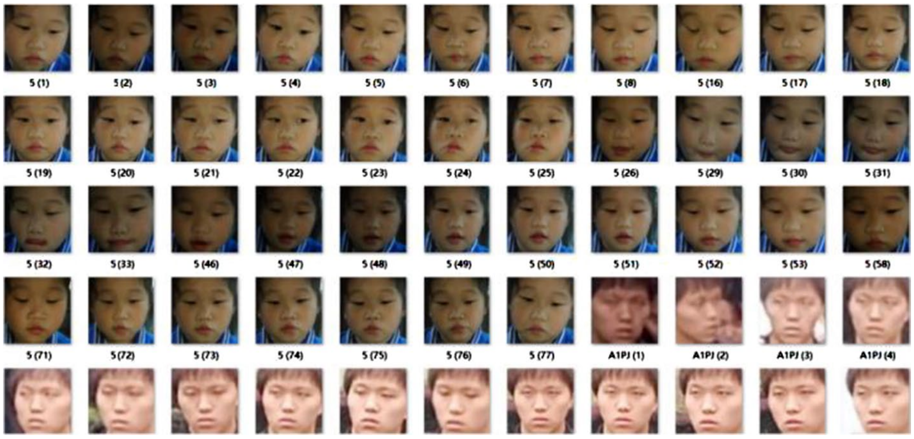


Fig. 7 The training set of the expression recognition model

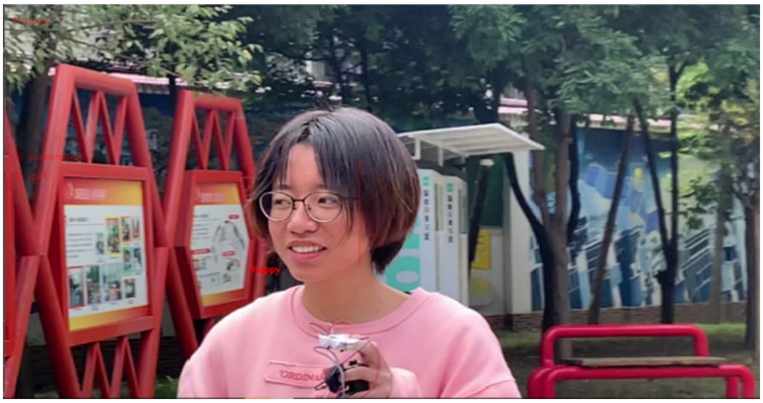


Fig. 8 Partial results of face detection and the expression recognition model

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).

$$\frac{\partial \text{Loss}}{\partial x_i} = \sum_i \frac{\partial \text{Loss}}{\partial y_i} \times \frac{\partial y_i}{\partial x_i}. \tag{6}$$

where y_i can be expressed by (7).

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

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

$$\frac{\partial y_i}{\partial x_j} = C_{ij}. \tag{8}$$

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

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

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

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

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 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 calculation of W' and H' after convolution.

$$W = S\left(W' - 1\right) - 2P + F, \tag{11}$$

$$H = S\left(H' - 1\right) - 2P + F. \tag{12}$$

Therefore, the structure of the expression recognition model is shown in Fig. 6. 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 complete the training of the model. Figure 7 shows partial data set utilized to train the expression recognition model.

Therefore, values of the model parameter 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 frame images appear each 1 s in the testing video. Following recognition process is defined so that the label of the expressions can be accurately defined within 1 s.

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 [27]. Figure 8 shows the partial results of face detection and the expression recognition model.

Therefore, the recognition result of expression in each frame image can be obtained. Figure 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 expression recognition result of the image per second. The expression in 1 s is considered to be 'Unhappy' if the number of 'Unhappy' frame images in 1 s is greater than 6 and greater than the number of 'Happy' frame images. If not, the judgments of other expressions are subsequently continued. The expression in 1 s is considered to be 'Happy' if the number of 'Happy' frame images in 1 s is greater than 4 and greater than the number of 'Unhappy' frame images. And the expression in 1 s is considered to be 'Happy' if the number of 'Happy' frame images in 1 s is greater than 0 and greater than the number of 'Unhappy' frame images when the number of 'Quiet' frame images in 1 s is less than 5. The expression per second is defined as 'Quiet' when 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. Above three numbers represent three emotions defined in Sect. 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 $E$
if $N_1 < 3$ or $N_2 < 3$ then
  The subjects have no emotional changes in this test
if $n_1 = 0$ and $n_2 > 1$ then
  $E$ = unhappy
if $n_1 > 5$ or '1' occurs four times in a row per 15s then
  $E$= happy
if $N_1 < 70$ and $n_{i1} > 2$ and $d_{11} < d_{12}$ then
  $n_1 = n_1$ + former $n_{i1}$
if $n_1 > 7$ then
  $E$= happy
else if $N_1 < 70$ and $n_{i1} > 2$ and $d_{11} > d_{12}$ then
  $n_1 = n_1$ + latter $n_{i1}$
  if $n_1 > 7$ then
    $E$ = happy
if $N_1 < N$ and $N_1 > \frac{1}{5}N$ and $n_1 > 4$ then
  $E$ = happy
else if $n_{i1} > 2$ and $d_{11} < d_{12}$ then
  $n_1 = n_1$ + former $n_{i1}$
  if $n_1 > 7$ then
    $E$= happy
else if $n_{i1} >2$ and $d_{11} > d_{12}$ then
  $n_1 = n_1$ + latter $n_{i1}$
  if $n_1 > 7$ then
    $E$= happy
if $N_1 < \frac{1}{5}N$ and $N_1 > \frac{1}{6}N$ and $n_1 > 3$ then
  $E$= happy
if $N_1 < \frac{1}{6}N$ and $N_1 > 12$ and $n_1 > 2$ then
  $E$= happy
if $N_2 < \frac{1}{6}N$ and $N_2 > 4$ and $n_2 > 1$ then
  $E$= unhappy
if $n_1 > 0$ and $n_2 > 0$ and $n_1 > n_2$ then
  $\mathrm{n}_1 = \mathrm{n}_1 + \mathrm{n}_2$
```

Where N represents the total number of expressions per second in the video. N_0, N_1, N_2 respectively indicate the number of corresponding expressions. n_0, n_1, n_2 are utilized to represent the number of different expressions in the video each 15 s. The number of 'Happy' and 'Unhappy' is defined as n_{i1} and n_{i2} of interval video in each 5 s to reduce the

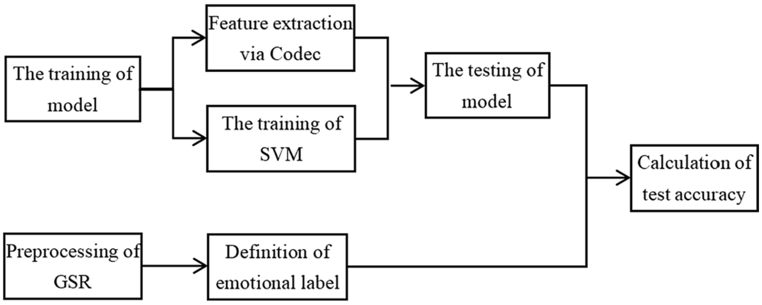

The training of model
Feature extraction via Codec
The training of SVM
The testing of model
Calculation of test accuracy
Preprocessing of GSR
Definition of emotional label

Fig. 9 The overall structure of the emotion recognition model

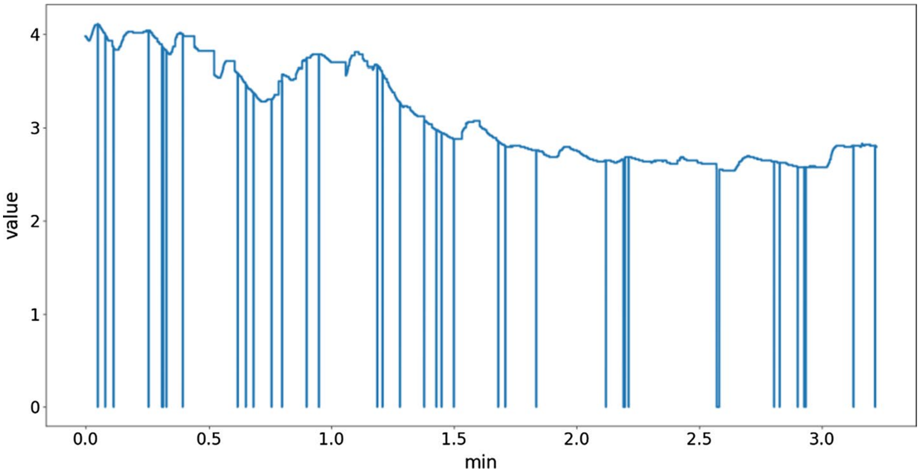


Fig. 10 The value of GSR during experiment for a subject

loss of useful information. In addition, the distance between first ‘1’ appearing in each interval video and last ‘1’ in 15 s video of previous section is defined as d_{11}. The distance between last ‘1’ appearing in each interval video and first ‘1’ in 15 s 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.

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. First, video needed to be grayed and standardized, then corresponding GSR needed to be completed data noise reduction and smoothing. Partial results of the latter was shown in Figs. 10 and 11.

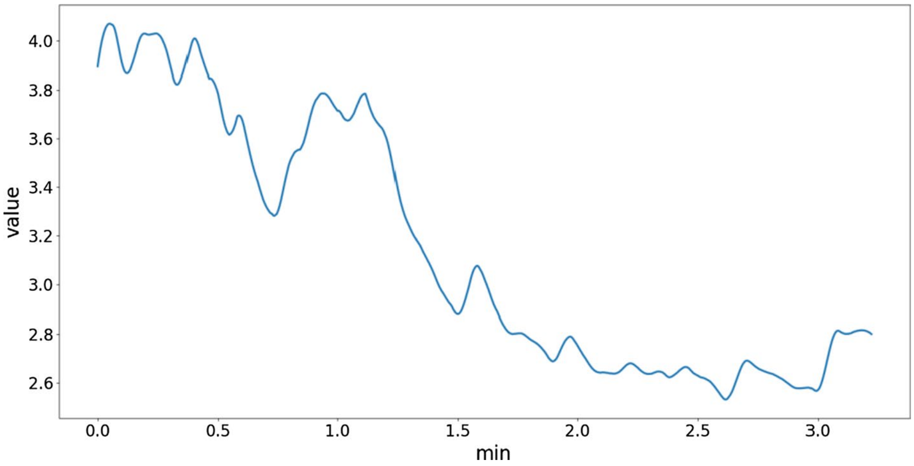


Fig. 11 The value of GSR after preprocessing for a subject

3.2 The setting of parameter

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

These parameter were 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 the comparison model.

The parameters of CascadeClassifier were shown in Table 2.

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.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}}. \tag{13}$$

Table 1 Training parameters

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

Table 2 Parameters of CascadeClassifier

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

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. Figure 12 showed the training part of the 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, but 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.

SVM classifier needed to be trained 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.

As shown in Fig. 14, it can be found that the highest accuracy was 97.71% when the kernel function was 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.

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.

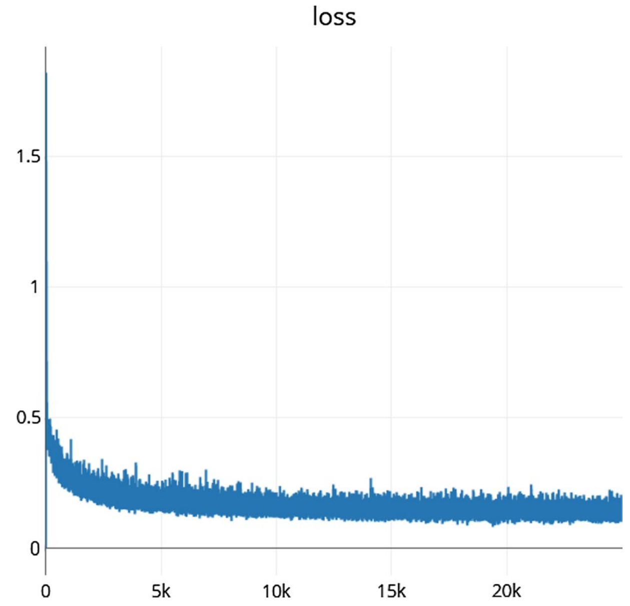


Fig. 12 The training process of the proposed model

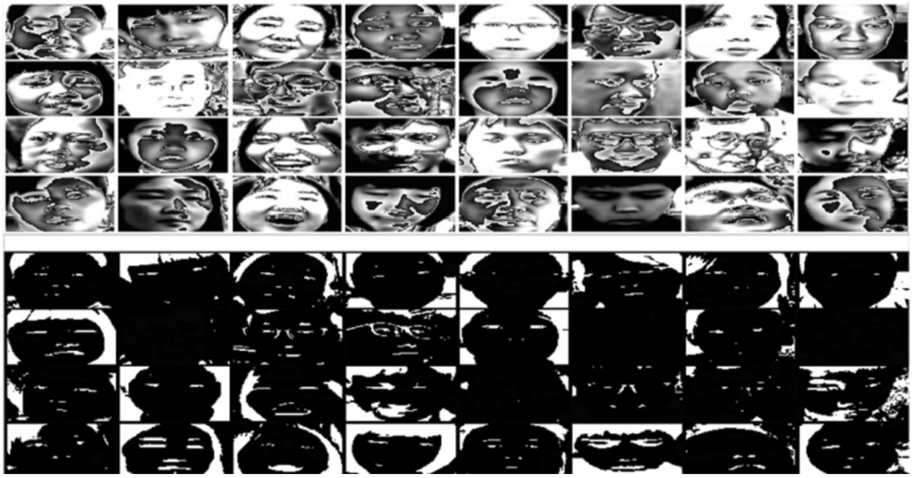


Fig. 13 Original image and generated image from its features

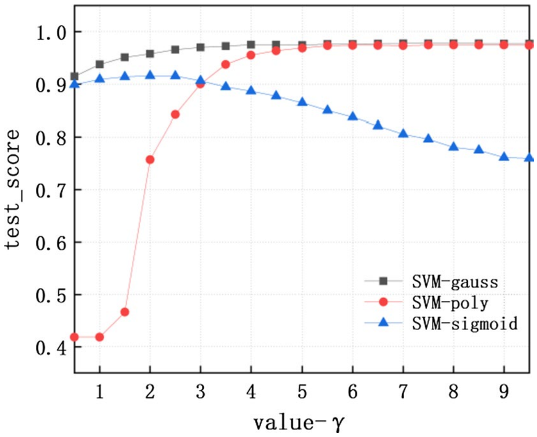


Fig. 14 The classification accuracy of SVM on testing set with different kernel and value of γ

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

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.

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

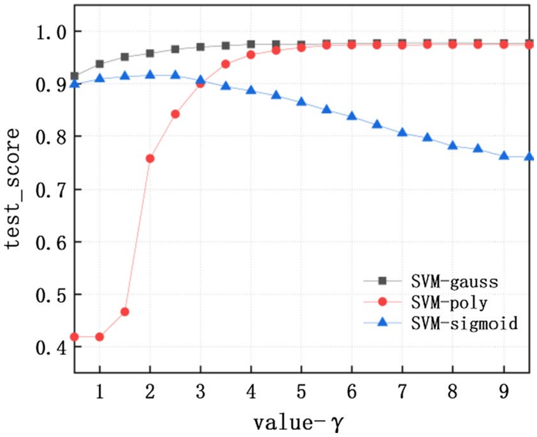


Fig. 15 The accuracy of classification changed with the value of *c*

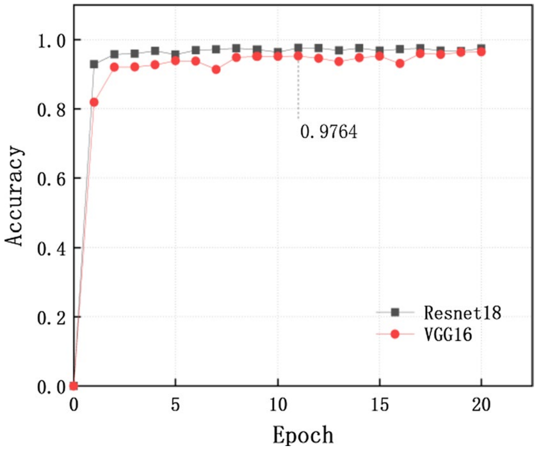


Fig. 16 The training process of Resnet18 and VGG16

proposed emotion recognition model to analyze the limitation of model. The matrix was shown in Fig. 18.

Following conclusions can be obtained by combing with the detailed result of emotion recognition in Figs. 17 and 18.

- The differences of facial styles leaded to the difficulty of facial emotion recognition in some subjects with unclear facial expressions or rich facial expressions. This class of subjects on 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 also leaded to the decrease of accuracy;
- The recognition of 'Quiet' emotion in 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.

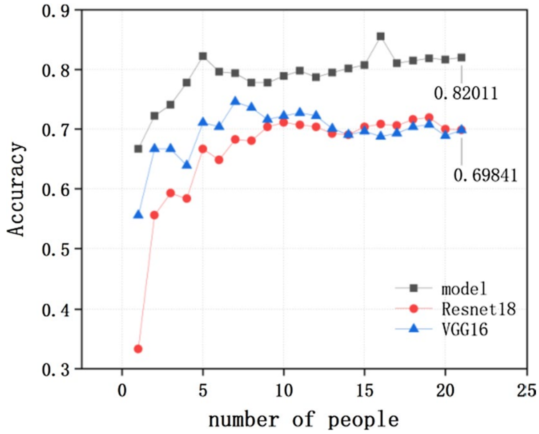


Fig. 17 The diagram of testing accuracy

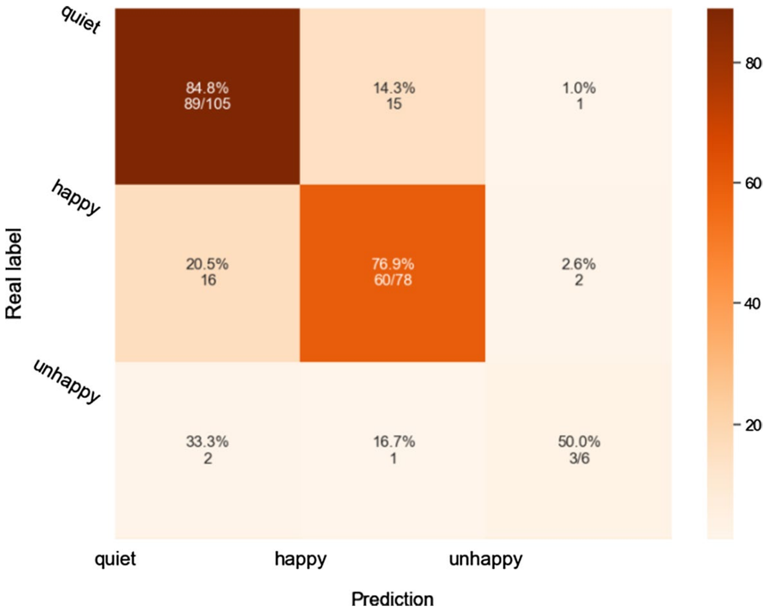


Fig. 18 Confusion matrix of the proposed emotion recognition model

4 Conclusions

Firstly, reliable emotional labels 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.

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.

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' 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 read and approved the final manuscript.

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).

Availability of data and materials

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

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

[1]College of Landscape Architecture, Sichuan Agricultural University, Chengdu 611130, China. [2]Guangdong Provincial Key Laboratory of Optical Information Materials and Technology and Institute of Electronic Paper Displays, South China Academy of Advanced Optoelectronics, South China Normal University, Guangzhou 510006, China.

Received: 10 August 2021 Accepted: 2 September 2021
Published online: 17 September 2021

References

1. X. Liu, X.P. Zhai, W.D. Lu, C. Wu, QoS-guarantee resource allocation for multibeam satellite industrial Internet of things with NOMA. IEEE Trans. Ind. Inform. **17**(3), 2052–2061 (2021). https://doi.org/10.1109/TII.2019.2951728
2. S.S. Devi, A. Bhuvaneswari, Quantile regressive fish swarm optimized deep convolutional neural learning for reliable data transmission in IoV. Int. J. Comput. Netw. Commun. **13**(2), 81–97 (2021). https://doi.org/10.5121/ijcnc.2021.13205
3. F. Valocky, M. Orgon, I. Fujdiak, Experimental autonomous car model with safety sensor in wireless network. IFAC PapersOnLine. **52**(27), 92–97 (2019). https://doi.org/10.1016/j.ifacol.2019.12.739
4. K. Afzal, R. Tariq, F. Aadil, Z. Iqbal, M. Sajid, An optimized and efficient routing protocol application for IoV. Math. Probl. Eng. (2021). https://doi.org/10.1155/2021/9977252
5. S. Turabzadeh, H.Y. Meng, R.M. Swash, M. Pleva, J. Juhar, Facial expression emotion detection for real-time embedded system. Technologies **6**, 17 (2018). https://doi.org/10.3390/technologies6010017
6. R.W. Picard, *Affective Computing: Challenges* (MIT Press, USA, 1997), pp. 2–10
7. K. Anderson, P.W. Mcowan, A real-time automated system for the recognition of human facial expressions. IEEE Trans. Cybern. **36**, 96–105 (2006). https://doi.org/10.1109/TSMCB.2005.854502
8. J. Ang, R. Dhillon, A. Krupski, E. Shriberg, A. Stolcke, Prosody-based automatic detection of annoyance and frustration in human–computer dialog, in *Seventh International Conference on Spoken Language Processing* (2002). p. 2037–2040.
9. C Feichtenhofer, A Pinz, A Zisserman, Convolutional Two-Stream Network Fusion for Video Action Recognition, in *Computer Vision and Pattern Recognition* (IEEE, 2016). p. 1933–1941. https://doi.org/10.1109/CVPR.2016.213
10. A. Mehrabian, Communication without words. Psychol. Today. **2**, 53–55 (1968). https://doi.org/10.1016/S0140-6736(65)90194-7
11. R.W. Picard, E. Vyzas, J. Healey, Toward machine emotional intelligence: analysis of affective physiological state. IEEE Trans. Pattern Anal. Mach. Intell. **23**, 1175–1191 (2016). https://doi.org/10.1109/34.954607
12. X. Liu, X.Y. Zhang, NOMA-based resource allocation for cluster-based cognitive industrial Internet of things. IEEE Trans. Ind. Inform. **16**, 5379–5388 (2020). https://doi.org/10.1109/TII.2019.2947435
13. N. Samadiani, G. Huang, W. Luo, C.H. Chi, Y.F. Shu, R. Wang, T. Kocaturk, A multiple feature fusion framework for video emotion recognition in the wild. Concurr. Comput. Pract. Exp.. (2020). https://doi.org/10.1002/cpe.5764
14. 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). https://doi.org/10.17660/ActaHortic.2005.671.18
15. S. Koelstra, C. Muhl, M. Soleymani, J.S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, Deap: a database for emotion analysis; using physiological signals. IEEE Trans. Affect. Comput. **3**, 18–31 (2012). https://doi.org/10.1109/T-AFFC.2011.15

16. M. Soleymani, J. Lichtenauer, T. Pun, M. Pantic, A multimodal database for affect recognition and implicit tagging. IEEE Trans. Affect. Comput. **3**, 42–55 (2012). https://doi.org/10.1109/T-AFFC.2011.25
17. W.R. Hu, G. Huang, L.L. Li, L. Zhang, Z.G. Zhang, Z. Liang, Video-triggered EEG-emotion public databases and current methods: a survey. Brain Sci. Adv. **6**, 255–287 (2019). https://doi.org/10.26599/BSA.2020.9050026
18. X. Liu, X.Y. Zhang, Rate and energy efficiency improvements for 5G-based IoT with simultaneous transfer. IEEE Internet Things J. **6**(4), 5971–5980 (2019). https://doi.org/10.1109/jiot.2018.2863267
19. X. Liu, X.Y. Zhang, M. Jia, L. Fan, W. Lu, X. Zhai, 5G-based green broadband communication system design with simultaneous wireless information and power transfer. Phys. Commun. **28**, 130–137 (2018). https://doi.org/10.1016/j.phycom.2018.03.015
20. P. Lucey, J.F. 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, in *Proceedings of the Third International Workshop on CVPR for Human Communicative Behavior Analysis(CVPR4HB)* (2010). p. 94–101. https://doi.org/10.1109/CVPRW.2010.5543262
21. C. Tronstad, H. Kalvøy, S. Grimnes, G. Martinsen-Ørjan, Improved estimation of sweating based on electrical properties of skin. Ann. Biomed. Eng. **41**, 1074–1083 (2013). https://doi.org/10.1007/s10439-013-0743-4
22. M.M. Bradley, P.J. Lang, Measuring emotion: behavior, feeling, and physiology, in *Cognitive Neuroscience of Emotion*, ed. by R.D. Lane, L. Nadel (Oxford University Press, New York, 2000). p. 242–276
23. P.J. Lang, Emotion and motivation: attention, perception, and action. J. Sport Exerc. Psychol. **22**, 180–199 (2020). https://doi.org/10.1097/00005131-200006000-00017
24. K.H. Kim, S.W. Bang, S.R. Kim, Emotion recognition system using short-term monitoring of physiological signals. Med. Biol. Eng. Comput. **42**, 419–427 (2004). https://doi.org/10.1007/BF02344719
25. H.H. Madden, Comments on Savitzky–Golay convolution method for least-squares fit smoothing and differentiation of digital data. Anal. Chem. **50**, 1383–1386 (1978). https://doi.org/10.1021/ac50031a048
26. V. Dumoulin, F. Visin, A Guide to Convolution Arithmetic for Deep Learning (2019), pp. 1–28
27. R. Lienhart, A. Kuranov, V. Pisarevsky, Empirical analysis of detection cascades of boosted classifers for rapid object detection, in *Joint Pattern Recognition Symposium* vol. 2781 (2003). p. 297–304. https://doi.org/10.1007/978-3-540-45243-0_39

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.