Hindawi
Scientific Programming
Volume 2022, Article ID 6315063, 11 pages
https://doi.org/10.1155/2022/6315063

Research Article

An Intelligent Vehicle Alarm User Terminal System Based on Emotional Identification Technology

Liping Wu,[1] **Maomao Liu**,[1] **Jingpeng Li**,[2] **and Yanjun Zhang**[1]

[1]*College of Information Science and Engineering, Shandong Agriculture and Engineering University, Jinan 250000, China*
[2]*Inspur Software Group Co., Ltd., Data Center Service Business Dept, Jinan 250000, China*

Correspondence should be addressed to Liping Wu; z2013463@sdaeu.edu.cn

Received 19 November 2021; Revised 21 December 2021; Accepted 30 December 2021; Published 14 February 2022

Academic Editor: Ahmed Farouk

Negative emotions could increase the risks of traffic accidents. However, the driver's emotional identification is rarely considered in the current design of intelligent vehicle alarm user terminals (IVAUTs). To solve the problem, this paper tries to design an IVAUT system based on emotional identification technology. Firstly, the transformer network was combined with a convolutional neural network (CNN) into a voice emotional identification system for intelligent vehicle alarm, and an emotional labeling approach was provided. Next, a bimodal fusion model was developed based on decompose-CNNs, which includes an appearance module, an optical flow module, and a bimodal fusion module. The proposed emotional identification approach was proved effective through experiments.

1. Introduction

Negative emotions could increase the risks of traffic accidents [1–8]. To assist with driving, it is necessary to effectively detect the driver's negative emotions and thus enhance his/her thinking ability, perception ability, and judgement ability [9–11]. Currently, emotions are mostly identified based on facial expressions, voices, postures, behaviors, texts, and physiological signals [12–18]. Among them, the identification technologies of voice signals, facial expressions, and physical behaviors are research hotspots. With the popularity of intelligent vehicle interaction equipment, emotional identification technology has been gradually introduced to assist with driving by monitoring fatigue driving and driver emotions [19, 20].

To prevent driving risks, Ooi et al. [21] proposed a new driver monitoring system. Specifically, a deep convolutional neural network (DCNN) was designed to recognize driver emotions, and an on-demand audio mechanism was developed to automatically collect audio resources with an online crawler, aiming to eliminate the driving risks induced by the driver's negative emotions. Based on a survey on electrodermal activities, Bi and Shen [22] recognized stress and anger as the main driver emotions that lead to accidents and develop a simulated driving operation with preset neutral, stress, and anger scenarios, according to emotional stimuli. Xie et al. [23] presented an emotion-based fatigue driving recognition algorithm for the drivers, who have been driving for a long time or have a poor mental state, and relied on the algorithm to prevent continuous fatigue driving, thereby avoiding incidents. Drawing on the Ortony-Clore-Collins (OCC) model of emotion, the Markov model of the automatic state transition of emotions, and the hidden Markov model of the state transition of emotions under stimuli, Neerincx et al. [24] modeled the driver emotions under two different road conditions and fixed road conditions and examined the variation in driver emotions during car following, lane changing, and overtaking. Riaz et al. [25] categorized the causes of driving emotions into two classes, namely, personal factors and specific driving conditions, pointing out that the nature and intensity of perceived emotions depend on the various evaluation factors under traffic conditions. Izquierdo-Reyes et al. [26] put forward an effective driver assistance model, which drives cognition by emotions. As an accident prevention scheme, the model considers the distraction of different types of drivers simultaneously.

After sorting out the domestic and foreign research, the studies on intelligent vehicle alarm user terminals (IVAUTs) focus on the design of terminals based on human-computer interaction and the development of information exchange interfaces based on design psychology. However, little attention has been paid to the emotional recognition of drivers in the IVAUT design. The proposed IVAUT system realizes the relevant functions based on techniques of voice emotion identification and facial emotion identification. The two emotion identification techniques were innovatively integrated to recognize and warn the negative emotions of drivers.

The available techniques of voice emotion identification cannot effectively recognize the individual difference in voice, while the current methods of facial emotion identification overlook the correlation between appearance modal and optical flow modal. Hence, this paper designs an innovative approach for the two identification technologies. Section 2 combines the transformer network with a convolutional neural network (CNN) into a voice emotional identification system for intelligent vehicle alarm and explains the emotional labeling approach. Section 3 develops a bimodal fusion model based on decompose-CNNs, which includes an appearance module, an optical flow module, and a bimodal fusion module. Section 4 verifies the effectiveness of our emotional identification approach through experiments.

2. Voice Emotion Identification System

With the technical advancement of various sensors and monitoring equipment, it is increasingly easier to acquire human voice signals and facial expression images with high accuracy. To ensure safe driving, it is particularly important to identify and warn drivers' negative emotions with artificial intelligence and machine learning algorithms, as well as the techniques of voice emotion identification and facial emotion identification. The IVAUTs can precisely monitor the real-time emotions of drivers, accurately grasp their psychological changes, and take timely countermeasures to prevent traffic accidents induced by negative emotions.

The voice features (e.g., pitch, tone, and loudness) of humans vary with emotional states. The existing techniques of voice emotion identification usually analyze the voice features corresponding to the known types of emotions, adjust the parameters and weights of the emotion identification model, making the model more effective in identifying emotions, and test the adjusted model.

This paper combines the transformer network with CNN to acquire voice emotional features more effectively. The primary voice emotional features were taken as the parameters of the two networks and transmitted them to the deep network. The selected CNN consists of four convolutional modules that optimize loss classifiers, a batch normalization module for regularization, a max-pooling layer that reduces the dimensionality of the feature map, a drop-out layer that prevents overfitting in training, and a rectified linear unit (ReLU) function to activate the standard layers. In the transformer network, the multihead attention encoder is connected to the fully connected feedforward network, which is followed by a ReLU function layer. The outputs of the transformer network and the CNN are combined and mapped by the softmax function to eight emotions.

2.1. Network Design. The transformer encoder structure was adopted to increase the number of features. If the data are batch-processed by instance normalization, the features extracted by the transformer network will differ slightly from those extracted by the CNN in terms of the internal relationship.

Moreover, the voice sample set for the IVAUTs covers the samples of 30 users. Since each user has unique voice features, the extracted features contain lots of personal features: the intraclass distance is even greater than the interclass distance. If the voice emotional features are classified by the softmax function, the intraclass distance would be enlarged, undermining the recognition effect of the model.

To solve the above problem, the voice emotion identification model with loss function was introduced to reduce the intraclass distance. The loss function learns the center of deep voice features in each class and penalizes the central features of the expected classes. In this way, the interclass distance is increased, and the intraclass distance is reduced. As shown in Figure 1, the established network consists of an input layer, multiple convolutional layers, a multihead attention mechanism, a fully connected layer, and an output layer.

The voice emotion model has a special requirement on the overall distribution of sample data. To keep the data distribution consistent, the batch normalization module in the network normalizes each batch of voice samples. The mean and variance of each batch are greatly affected by the data size of that batch. If the data are too few, it is impossible to characterize the sample distribution with the calculated mean and variance. Voice emotion recognition usually deals with the entire input sentence. Let F be the number of hidden units on each layer; k be the k-th hidden layer; v be the value of a node before activation. To obtain more, richer emotional features of voice sentences, layer normalization was adopted in our transformer encoder. The input of all nodes is regularized by

$$N^k = \frac{1}{F}\sum_{i=1}^{F} v_i^k, \tag{1}$$

$$\varepsilon^k = \sqrt{\frac{1}{F}\sum_{i=1}^{F}\left(v_i^k - N^k\right)^2}. \tag{2}$$

By formulas (1) and (2), the expectation N and standard error ε can be obtained for each layer. The hidden units in the same layer share the same normalized expectation N and standard error ε. However, the networks trained on different cases would have different normalized expectations N and

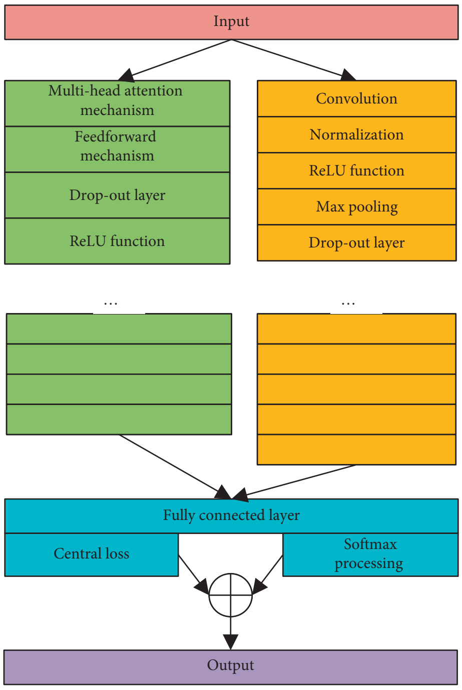


Figure 1: Structure of our voice emotion identification network.

standard error ε. Let h be the gain and r be the offset. Then, we have

$$v_i^{-k} = \frac{h^k}{\varepsilon^k} \cdot \left(v_i^k - N^k\right) + r. \tag{3}$$

This paper classifies the voice emotional features with softmax cross-entropy loss function. Identical to the m-classification problem, the inputs of our network can be mapped to real numbers in [0, 1] through regularization, with the sum of all inputs being 1. For any voice sample E, the probability of being assigned to voice emotion class D_i is GL_i. Then, we have

$$GL_1 + GL_2 + \cdots + GL_m = 1. \tag{4}$$

The voice emotion class of voice sample E can be expressed as

$$\max\left(GL_1, GL_2, \ldots, GL_m\right). \tag{5}$$

According to the requirements of the m-classification problem, the softmax function outputs the probability GL_i of each voice emotion class D_i. The maximum probability max (GL_1, GL_2, ...GL_m) is the most probable class of the inputs. The inputs of the softmax function consist of transformer network output and CNN output. Let δ and γ be the hyperparameters to be adjusted during the training; ω_{ij} be the j-th weight of node i related to the CNN output; ω_{nm} be the n-th weight of node m related to transformer network output. Then, the i-th output can be expressed as

$$C_i = \sum_j \omega_{ij}^C + r^C + \gamma\left(\sum_m \omega_{nm}^{TR} + r^{TR}\right). \tag{6}$$

Next, the softmax function is added to the network output. Then, the i-th output O_k can be expressed as

$$O_i = \frac{e^{C_i}}{\sum_{j=1}^m e^{C_i}}. \tag{7}$$

The cross-entropy function, which is easy to derive and compute, was selected as the softmax loss function [24]. Let b_i be the actual classification result. The cross-entropy function can speed up the network learning:

$$L = -\sum_i b_i \ln v_i. \tag{8}$$

2.2. Loss Function. To facilitate the design of the classifier for voice emotion identification, the distance between different classes of voice emotions should be maximized, while that between the same class of voice emotions under different scenarios should be minimized. As a common classification method for multiclass tasks, the softmax cross-entropy loss function can learn the separable features in order to differentiate between various voice emotions. For this purpose, the loss function was introduced to the proposed voice emotion identification model, which draws voice emotions to the center of their corresponding voice emotion classes. Let d_j be the center of the i-th class of voice emotions. Then, the loss function can be expressed as

$$L_{\text{Center}}^* = -\frac{1}{n}\sum_{i=1}^n \left\|c_i - d_{b_i}\right\|^2. \tag{9}$$

By optimizing the loss function, the distance between the same class of voice emotional features will become smaller.

After initializing d_j as 0 and defining the class center of each minibatch as $\dot{d}_j$, the class center d'_j of the j-th type of features in the minibatch can be calculated by

$$d'_j = \frac{\sum_{i=1}^{n} \xi (b_i = j) C_i}{\sum_{i=1}^{n} \xi (b_i = j) > 0}. \tag{10}$$

If $b_i = j$, then $\xi\ (b_i = j) C_i = 1$. Let δ be the hyperparameter that controls the update rate of d_j, if the new minibatch has the voice emotional features corresponding to the j-th class of emotions; d_j^{τ} and $\dot{d}_j^{\tau}$ be the values of d_j and $\dot{d}_j$ in the τ-th iteration, respectively. In other cases, $\xi\ (b_i = j) C_i = 0$. Then, the class center d_j can be defined as

$$d_j^{\tau+1} = \begin{cases} (1-\delta) d_j^{\tau} + \delta \dot{d}_j^{\tau}, & \sum_{i=1}^{n} \xi (b_i = j) > 0, \\ d_j^{\tau}, & \sum_{i=1}^{n} \xi (b_i = j) = 0. \end{cases} \tag{11}$$

Let θ_j be the inverse ratio of class j voice emotions in the entire voice sample training set. Because of the imbalance between voice emotion classes, two weights were assigned to softmax cross-entropy loss function and loss function:

$$\begin{aligned} L_{\text{SoftMax}} &= \frac{1}{\sum_{i=1}^{n} \theta_{b_i}} \sum_{i=1}^{n} \theta_{b_i} \ln v_i, \\ L_{\text{Center}} &= \frac{1}{\sum_{i=1}^{n} \theta_{b_i}} \left\| c_i - d_{b_i} \right\|^2. \end{aligned} \tag{12}$$

Hence, this paper uses a joint loss function to train the proposed voice emotion identification model. Let μ be the hyperparameter that balances central loss with softmax cross-entropy loss. Then, the joint loss function composed of softmax cross-entropy loss function and loss function can be expressed as

$$L = L_{\text{SoftMax}} + \mu L_{\text{Center}}. \tag{13}$$

2.3. Emotional Labeling. The proposed voice emotion identification model was trained on the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), which contains eight types of emotions with a frame precision of nearly 80%. The voice samples were labeled directly by the naming rule of voice emotion data in that database. Each voice sample was named in the format of "audio number-data type-emotion label-data intensity-semantic content-repetition-user number.wav." Specifically, the data type is speech 01 or song 02; the emotion label could be one of the following: indifferent 01, calm 02, happy 03, sad 04, angry 05, fearful 06, disgust 07, and surprise 08; the data intensity is either normal 01 or strong 02; the repetition is 01 (first repetition) or 02 (second repetition).

In the RAVDESS, the emotions are classified into eight types 01–08, which are common in most datasets. To ensure the emotional recognition effect of our model on subjects of different voice features and maintain the diversity of voice emotions, this paper randomly divides the voice sample set into eight subsets. Five of them were used for training and the remaining three for testing.

3. Facial Emotion Identification System

Each facial expression is a dynamic process in a period of time, which is induced by the movement of facial muscles. Compared with the single frame-based static facial expression identification, the dynamic identification oriented at vehicle videos can acquire facial expression features with high accuracy. This paper sets up a bimodal fusion model based on decompose-CNNs (Figure 2), which mainly consists of an appearance module, an optical flow module, and a bimodal fusion module. Based on three-dimensional (3D) decompose-CNN, the appearance module processes the red-green-blue (RGB) image sequence of video frames. Based on two-dimensional (2D) decompose-CNN, the optical flow module processes a single optical flow image, which normally encompasses a start frame and vertex frames. The bimodal fusion module applies a consistency constraint on the emotion labels predicted by facial expressions of different modals, such as integrating the information of the appearance module with that of the optical flow module in the feature space.

The time information and spatial information of a frame can be extracted by one-dimensional (1D) convolution and 2D convolution, respectively. 1D convolution, which operates in the height and width directions, can be naturally decomposed through 2D convolution. Both 1D and 2D convolutions can be decomposed by 3D convolution. Let Γ(.) be a complete 2D convolutional layer; Q^{φ} and Q^{ϕ} be kernels of the size $1 \times R$ and $R \times \mathrm{l}$, respectively; $\Gamma_{3D}(.)$ be a 3D convolutional layer; Q^{ρ} and Q^{σ} be the kernels of the size$1 \times R \times R$ and $R \times \mathrm{l} \times \mathrm{l}$, respectively; $\mathrm{P} \in \mathbb{R}^{S \times Q \times U\phi}$ and $P \in \mathbb{R}^{\Psi \times S \times Q \times U\sigma}$ be the final feature maps output by the decompose-CNNs, respectively; Ψ, S, and Q be the time length, height, and width, respectively; U^{ϕ} and U^{σ} be the number of channels. Then, this paper defines the two types of decompose-CNNs. The feature maps G and G output by the preceding network layer can be, respectively, calculated by the following decompose-convolutions:

$$\begin{aligned} \mathrm{P} &= \Gamma\left(Q^{\varphi}, \Gamma\left(Q^{\phi}, G\right)\right), \\ P &= \Gamma_{3D}\left(Q^{\sigma}, \Gamma_{3D}\left(Q^{\rho}, G\right)\right). \end{aligned} \tag{14}$$

The decomposition of 2D and 3D convolutions greatly reduces the number of parameters and computing complexity of the network, resulting in more efficient recognition of facial expressions. Besides, the network layers are increased to enhance the ability of the network to handle nonlinear data. Figure 3 shows the decomposition process of 2D convolution. The decomposition process of 3D convolution is similar to Figure 3.

The optical flow indicates the direction and intensity of pixel motion in video frames. Suppose pixel (a), (b) moves to $(a+\Delta a,\ b+\Delta b)$ from moment ψ to moment $\psi+\Delta\psi$. Let v and s be the horizontal and vertical components of the optical flow. Then, the vector of the optical flow can be calculated by

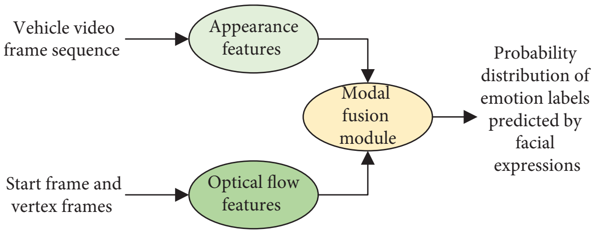

Vehicle video frame sequence
Appearance features
Modal fusion module
Probability distribution of emotion labels predicted by facial expressions
Start frame and vertex frames
Optical flow features

Figure 2: Architecture of bimodal fusion model.

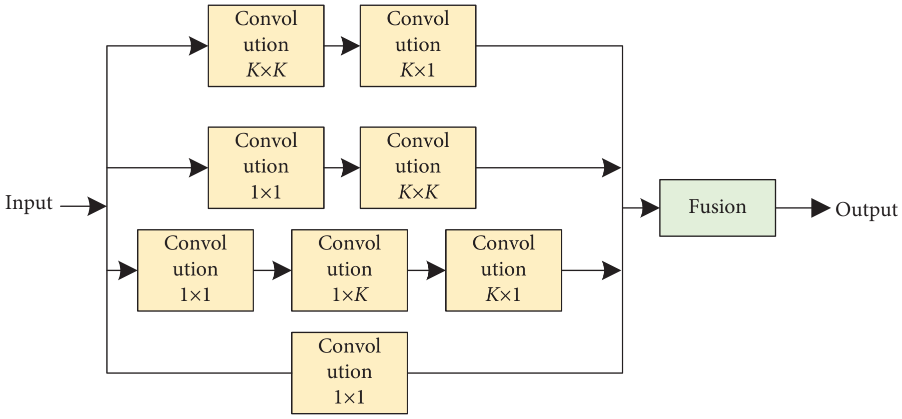


Figure 3: Decomposition process of 2D convolution.

$$q = (v, s)^T = \left(\frac{\Delta a}{\Delta \psi}, \frac{\Delta b}{\Delta \psi}\right)^T. \tag{15}$$

The motion state between frames can be described by the optical flow vector between the dense optical flow images of adjacent car video frames. If the car video contains k frames, it is possible to obtain k-1 optical flow images. The more the frames, the more complex the extraction of optical flow.

The facial expressions in car video frames can be perceived synergistically from different angles, for example, appearance information and optical flow information. Figure 4 shows the structure of 2D decompose-CNN. If the two modals of appearance features and optical flow features of facial expressions are simply connected in series, the correlation between the two kinds of features will be ignored. To fully utilize the correlation, this paper fuses the two modals consistently and constructs a classification constraint based on fused features, aiming to reduce the classification difference of facial expressions between the two modals.

Let χ_i be the eigenvector of the i-th sample; GL_1 and GL_2 be the probability distributions of the same length. Then, the matching degree of GL_2 to GL_1 can be measured by relative entropy Φ:

$$\Phi\left(GL_1 \| GL_2\right) = -GL_1\left(\chi_i\right)\log\frac{GL_1\left(\chi_i\right)}{GL_2\left(\chi_i\right)}. \tag{16}$$

Because of the asymmetry of relative entropy Φ, that is, $\Phi(GL_1||GL_2) \neq \Phi(GL_2||GL_1)$, the difference between the two probability distributions can be characterized by symmetric relative entropy Φ^*:

$$\Phi^*\left(GL_1 \| GL_2\right) = \frac{1}{2}\Phi\left(GL_1 \| GL_2\right) + \frac{1}{2}\Phi\left(GL_2 \| GL_1\right). \tag{17}$$

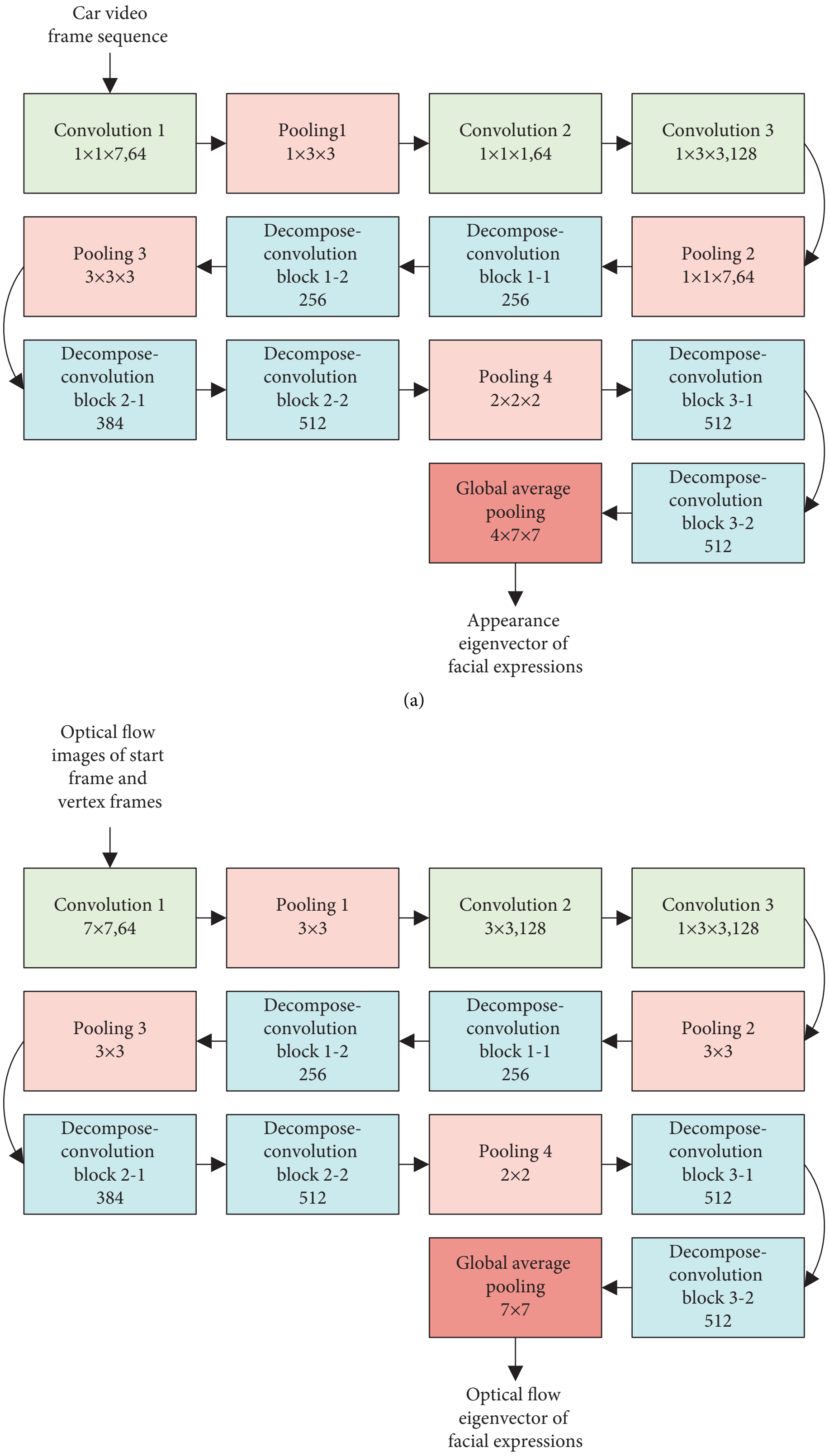
Let χ^t be the eigenvector obtained through the consistent fusion between appearance eigenvector χ^n and optical flow eigenvector χ^m; M be the number of samples; β_i be the probability distribution of the actual labels of the i-th sample; $GL(\chi^t{}_i)$ be the probability distribution of predicted labels obtained based on the fused eigenvector $\chi^t{}_i$ of the i-th sample. Then, the classification constraint can be calculated by

$$\Omega = \sum_{i\in M} \Phi^*\left(\beta_i \| GL\left(\chi_i^t\right)\right). \tag{18}$$

Let W be the number of classes; ζ_χ be the parameter vector of class w prediction. Then, $GL(\chi^t{}_i)$ can be calculated by softmax function:

$$GL\left(\chi_i\right) = \frac{1}{\sum_{w=1}^{W} e^{\zeta_a^T \chi_j}}\left[e^{\zeta_1^T \chi_i}, e^{\zeta_2^T \chi_i}, \ldots, e^{\zeta_w^T \chi_i}\right]^T. \tag{19}$$

In theory, the appearance features and optical flow features predicted for the same input should have similar modals; that is, the two types of features should be consistent. To prevent the extraction error of a single modal from affecting the model output, this paper applies a



(a)

(b)

Figure 4: Structure of 2D decompose-CNN.

consistency constraint on the probability distributions $GL(\chi_i^{\ n})$ and $GL(\chi_i^{\ m})$ of predicted feature labels corresponding to the two modals:

$$\text{PLPD} = \eta \sum_{i\in M} \Phi^{*}\left(GL\left(\chi_i^n\right) \| GL\left(\chi_i^m\right)\right), \tag{20}$$

where η is a hyperparameter. Formula (20) ensures the effective synthesis of complementary information of the two modals. Let $\Delta = \{\zeta^n, \zeta^m, \zeta^t\}$ be the online parameter of the modal fusion module. The final loss function can be given by

$$L(\Delta) = \Omega + \text{PLPD}. \tag{21}$$

Figure 5 explains the flow of modal fusion. Gradient descent was further adopted to optimize the network loss function, such that the probability distribution of predicted feature labels based on fused facial expressions approximates the probability distribution of actual feature labels. The partial derivative of the loss function can be solved by

$$\begin{aligned}\frac{\partial L(\Delta)}{\partial \zeta^n} &= \eta \sum_{i=1}^{M} \frac{\partial \Phi^{*}\left(GL\left(\chi_i^n\right) \| GL\left(\chi_i^m\right)\right)}{\partial \zeta^n}\\ &= \frac{\eta}{2} \sum_{i=1}^{M} \frac{\partial \Phi\left(GL\left(\chi_i^n\right) \| GL\left(\chi_i^m\right)\right)}{\partial \zeta_{jk}^n} + \frac{\partial \Phi\left(GL\left(\chi_i^m\right) \| GL\left(\chi_i^n\right)\right)}{\partial \zeta_{jk}^n}\\ &= -\frac{\eta}{2} \sum_{i=1}^{M} \left[\sum_{w=1}^{W} \left(\frac{GL\left(w|\chi_i^m\right)}{GL\left(w|\chi_i^n\right)} + \ln GL\left(w|\chi_i^m\right)\right) \frac{\partial GL\left(w|\chi_i^n\right)}{\partial \zeta_{jk}^n}\right],\end{aligned} \tag{22}$$

where

$$\begin{aligned}\frac{\partial GL\left(w|\chi_i^n\right)}{\partial \zeta_{jk}^n} &= \frac{\partial e^{\zeta_w^{nT} \chi_i^n} / \sum_{w=1}^{W} e^{\zeta_w^{nT} \chi_i^n}}{\partial \zeta_{jk}^n}\\ &= \left(\xi(w = j) - GL\left(w|\chi_i^n\right)\right) GL\left(w|\chi_i^n\right) \chi_{ik}^n.\end{aligned} \tag{23}$$

The probability distribution of labels predicted by appearance features approaches that of labels predicted by optical flow features, enhancing the robustness of the joint optimization and fused features. Hence, the emotion identification effect of our model could be improved. Let $\zeta^n{}_j$ be a subvector of ζ^n; $\zeta^n{}_{jk}$ be the k-th element in $\zeta^n{}_j$. The derivatives of the first and second terms of formula (21) can be solved similarly relative to ζ^m and ζ^t.

4. Experiments and Results Analysis

Hyperparameters δ, μ, and γ are the weight of class center update rate, weight of loss function, and weight of encoder output, respectively. This paper designs a contrastive experiment to explore the influence of these hyperparameters on the voice emotion recognition effect. The voice emotion recognition accuracy under different hyperparameter settings is reported in Figure 6, where features A and B are Mel cepstral coefficient (MCC) and Mel frequency cepstral coefficient (MFCC), respectively. Comparing the three subgraphs of Figure 6, the voice emotion recognition accuracy was not sensitive to δ but significantly affected by μ and γ. If μ and γ are too large or too small, the recognition accuracy would be very low.

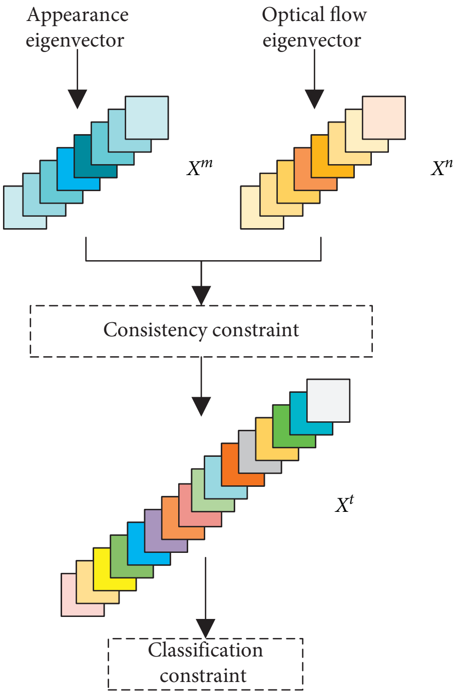


FIGURE 5: Flow of modal fusion.

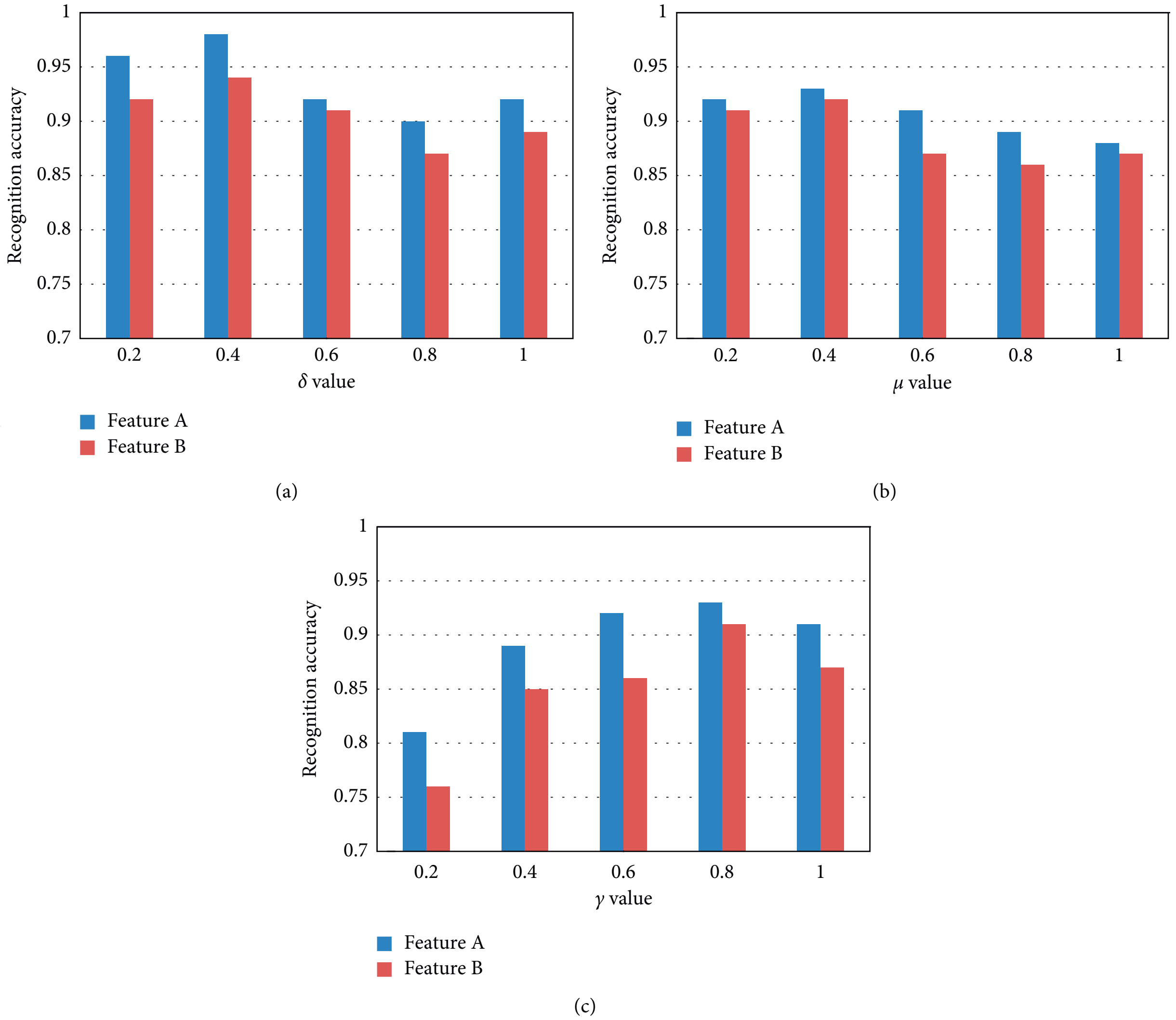
Figure 7 displays the confusion matrix of our voice emotion identification network based on MFCC. Our model performed excellent on the recognition of calm, sad, surprise, and disgust but made a few errors in recognition of the other four emotions.

Table 1 compares the emotion recognition rates of different models. Compared with traditional CNN, the combined network achieved relatively high accuracy in emotion recognition: the recognition rate surpassed 80% on all emotions, except calm (77.45%); the highest recognition rate was realized on disgust (91.72%). After introducing L_{center}, our model became much more accurate than the combined network. Therefore, the introduction of L_{center} can enhance the effectiveness of our model in voice emotion identification.

Furthermore, MCC was imported to our network. The hyperparameters were set to $\delta = 0.85$, $\mu = 0.15$, and $\gamma = 1$ after sensitivity analysis. Six experiments were carried out, and the most accurate results were selected for analysis. Figure 8 presents the MCC-based confusion matrix, and Table 2 lists the recognition rate of each voice emotion. Obviously, the classification effect of voice emotions could be improved by transformer network and loss function.

This paper compares the MCC- and MFCC-based emotion recognition scores of CNN, combined network, and our model. The results of the three models are compared in Table 3. The results show that our model coupled with MCC achieved the best results among all possible combinations. This is the best combination for emotion recognition of the selected voice emotion sample set.

Figure 9 compares the emotion recognition accuracies of facial expressions under different fusion mechanisms. Different fusion mechanisms had similar accuracy and

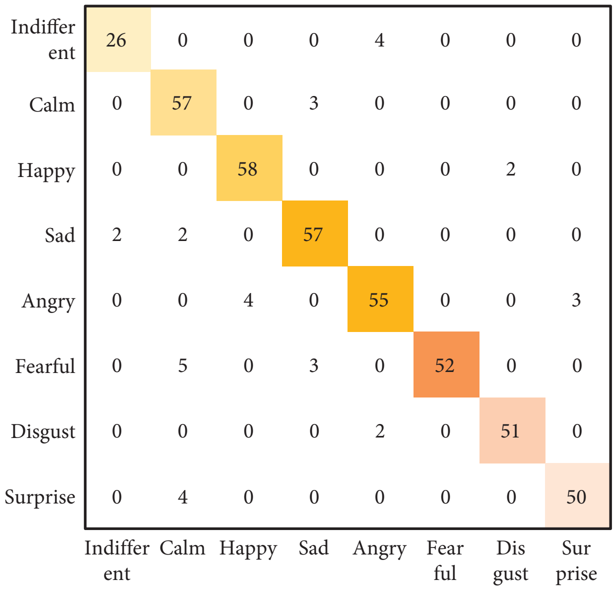


(a)

(b)

(c)

Figure 6: Voice emotion recognition accuracy under different hyperparameter settings.

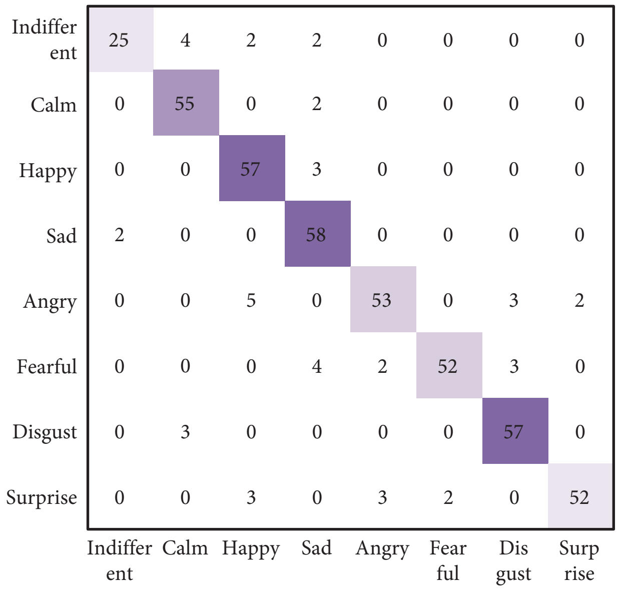


	Indifferent	Calm	Happy	Sad	Angry	Fearful	Disgust	Surprise
Indifferent	26	0	0	0	4	0	0	0
Calm	0	57	0	3	0	0	0	0
Happy	0	0	58	0	0	0	2	0
Sad	2	2	0	57	0	0	0	0
Angry	0	0	4	0	55	0	0	3
Fearful	0	5	0	3	0	52	0	0
Disgust	0	0	0	0	2	0	51	0
Surprise	0	4	0	0	0	0	0	50

Figure 7: MFCC-based confusion matrix.

Table 1: MFCC-based emotion recognition rates.

Name of model	CNN	Combined network (before introducing L_{center})	Our model (after introducing L_{center})
Indifferent (%)	83.46	85.79	87.35
Calm (%)	82.72	77.45	97.45
Happy (%)	82.35	88.72	86.54
Sad (%)	87.54	81.26	97.72
Angry (%)	87.36	84.54	88.46
Fearful (%)	88.41	88.46	88.72
Disgust (%)	83.58	91.72	93.27
Surprise (%)	83.21	90.21	95.31



	Indiffer ent	Calm	Happy	Sad	Angry	Fear ful	Dis gust	Surp rise
Indiffer ent	25	4	2	2	0	0	0	0
Calm	0	55	0	2	0	0	0	0
Happy	0	0	57	3	0	0	0	0
Sad	2	0	0	58	0	0	0	0
Angry	0	0	5	0	53	0	3	2
Fearful	0	0	0	4	2	52	3	0
Disgust	0	3	0	0	0	0	57	0
Surprise	0	0	3	0	3	2	0	52

Figure 8: MCC-based confusion matrix.

Table 2: MCC-based emotion recognition rates.

Name of model	CNN	Combined network (before introducing L_{center})	Our model (after introducing L_{center})
Indifferent (%)	87.22	91.26	94.26
Calm (%)	79.75	83.21	83.21
Happy (%)	82.35	90.72	96.72
Sad (%)	81.54	92.38	98.38
Angry (%)	84.26	88.46	88.46
Fearful (%)	85.58	89.72	89.72
Disgust (%)	87.21	90.21	96.21
Surprise (%)	89.54	82.34	92.34

Table 3: Performance scores of different models.

Name of model		Precision (%)	Recall (%)	F1-score (%)
MFCC	CNN	92.35	90.72	91.54
	Combined network (before introducing L_{center})	92.35	92.31	92.23
	Our model	94.23	94.20	94.20
MCC	CNN	91.62	91.23	91.76
	Combined network (before introducing L_{center})	93.57	93.62	93.85
	Our model	95.48	94.48	94.48

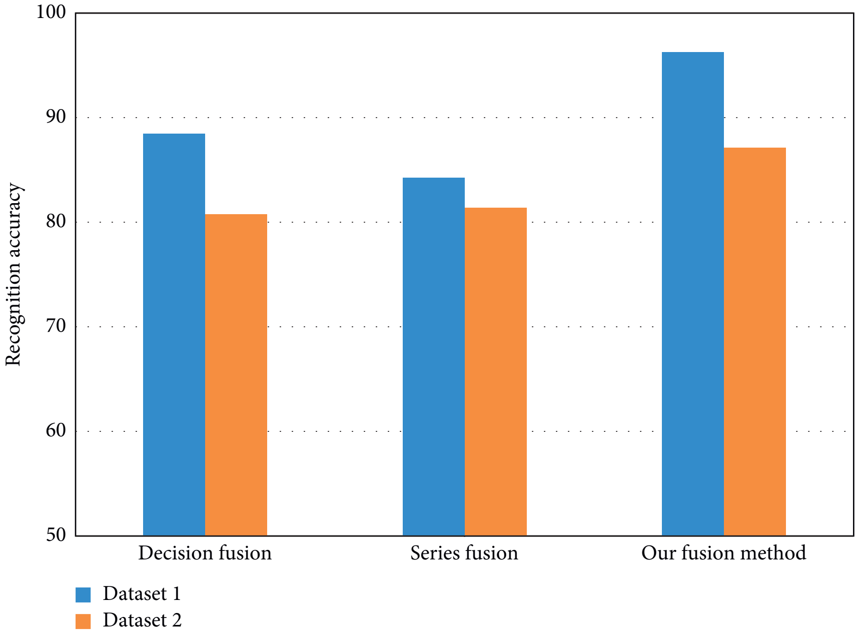


Figure 9: Emotion recognition accuracies of facial expressions under different fusion mechanisms.

precision trends on the two datasets, which involve different users. Our model achieved better recognition accuracy than decision fusion, which cannot characterize the distribution features of different features, and series fusion, which merely stitches up the information from different modals. The main reason is that our model further refines the loss function and thus shortens the distance between the predicted label distributions of different modals.

5. Conclusions

This paper mainly develops an IVAUT system based on emotion identification technology. Specifically, a combined network was designed to identify voice emotions for IVAUTs, based on the transformer network and the CNN, and the setting of emotion labels was explained in detail. Next, a bimodal fusion model was developed based on decompose-CNNs. There are three major modules in the model: appearance, optical flow, and bimodal fusion. After that, the voice emotion recognition accuracy was analyzed under different hyperparameter settings, the confusion matrices of our model were established based on MFCC and MCC, respectively, and the emotion recognition accuracies of different models on MFCC and MCC were calculated. The results show that our model, coupled with MCC, achieved better results than any other combination. Finally, the authors tested the emotion recognition accuracies of facial expressions under different fusion mechanisms and confirmed the effectiveness of our emotion identification approach.

The significance of this research lies in identifying the negative emotions of drivers in voices and facial images and making warning and intervention via IVAUTs, aiming to prevent emotional driving behaviors. The research provides a reference for the design and development of future smart transportation systems.

Due to the difficulty in acquiring lots of real data, this research faces several limitations: the conditions and means of emotional identification experiments are very limited, and the test and evaluation data were insufficiently quantified for the voice and facial emotion identification techniques applied in IVAUTs. Future work will try to overcome these limitations through in-depth research.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] T. E. Guerrero, J. de Dios Ortuzar, and S. Raveau, "Traffic accident risk perception among drivers: a latent variable approach," *Transportation Planning and Technology*, vol. 43, no. 3, pp. 313–324, 2020.

[2] H. T. Zhao, H. L. Cheng, Y. Ding, H. Zhang, and H. B. Zhu, "Research on traffic accident risk prediction algorithm of edge internet of vehicles based on deep learning," *Dianzi Yu Xinxi Xuebao/Journal of Electronics and Information Technology*, vol. 42, no. 1, pp. 50–57, 2020.

[3] A. Fang, C. Qiu, L. Zhao, and Y. Jin, "Driver risk assessment using traffic violation and accident data by machine learning approaches," in *Proceedings of the 2018 3rd IEEE International Conference on Intelligent Transportation Engineering (ICITE)*, pp. 291–295National University of Singapore, Singapore, September 2018.

[4] G. Waizman, S. Shoval, and I. Benenson, "Traffic accident risk assessment with dynamic microsimulation model using range-range rate graphs," *Accident Analysis & Prevention*, vol. 119, pp. 248–262, 2018.

[5] L. Eboli, G. Mazzulla, and G. Pungillo, "Measuring the driver's perception error in the traffic accident risk evaluation," *IET Intelligent Transport Systems*, vol. 11, no. 10, pp. 659–666, 2017.

[6] R. Ngueutsa and D. R. Kouabenan, "Accident history, risk perception and traffic safe behaviour," *Ergonomics*, vol. 60, no. 9, pp. 1273–1282, 2017.

[7] Y. Shiomi, K. Watanabe, H. Nakamura, and H. Akahane, "Assessing safety of signalized intersections: Influence of geometric attributes and regionality on traffic accident risk," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2659, no. 1, pp. 71–79, 2017.

[8] I. Norros, P. Kuusela, S. Innamaa, E. Pilli-Sihvola, and R. Rajamäki, "The Palm distribution of traffic conditions and its application to accident risk assessment," *Analytic methods in accident research*, vol. 12, pp. 48–65, 2016.

[9] A. Eherenfreund-Hager, O. Taubman – Ben-Ari, T. Toledo, and H. Farah, "The effect of positive and negative emotions on young drivers: A simulator study," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 49, pp. 236–243, 2017.

[10] Y. Zhu, Y. Wang, G. Li, and X. Guo, "Recognizing and releasing drivers' negative emotions by using music: evidence from driver anger," in *Adjunct Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pp. 173–178, New York, NY, USA, October 2016.

[11] M. Jeon, J. B. Yim, and B. N. Walker, "An angry driver is not the same as a fearful driver: effects of specific negative emotions on risk perception, driving performance, and workload," in *Proceedings of the 3rd international conference on automotive user interfaces and interactive vehicular applications*, pp. 137–142, New York, NY, USA, November 2011.

[12] X. Zhang, C. Xu, W. Xue, J. Hu, Y. He, and M. Gao, "Emotion recognition based on multichannel physiological signals with comprehensive nonlinear processing," *Sensors*, vol. 18, no. 11, p. 3886, 2018.

[13] K. P. Seng, L. M. Ang, and C. S. Ooi, "A combined rule-based and machine learning audio-visual emotion recognition approach," *IEEE Transactions on Affective Computing*, vol. 9, no. 1, pp. s3–13, 2016.

[14] N. Jain, S. Kumar, A. Kumar, P. Shamsolmoali, and M. Zareapoor, "Hybrid deep neural networks for face emotion recognition," *Pattern Recognition Letters*, vol. 115, pp. 101–106, 2018.

[15] A. Rajasekhar and M. K. Hota, "A study of speech, speaker and emotion recognition using Mel frequency cepstrum coefficients and support vector machines," in *Proceedings of the 2018 International Conference on Communication and Signal Processing (ICCSP)*, pp. 0114–0118, Chennai, India, April 2018.

[16] C. Maramis, L. Stefanopoulos, I. Chouvarda, and N. Maglaveras, "Emotion recognition from haptic touch on android device screens, precision medicine powered by phealth and connected health," in *Proceedings of the International Conference on Biomedical and Health Informatics*, pp. 205–209, Thessaloniki, Greece, November 2017.

[17] S. Klaylat, Z. Osman, L. Hamandi, and R. Zantout, "Emotion recognition in Arabic speech," *Analog Integrated Circuits and Signal Processing*, vol. 96, no. 2, pp. 337–351, 2018.

[18] Z.-T. Liu, Q. Xie, M. Wu, W.-H. Cao, Y. Mei, and J.-W. Mao, "Speech emotion recognition based on an improved brain emotion learning model," *Neurocomputing*, vol. 309, pp. 145–156, 2018.

[19] E. Roidl, B. Frehse, M. Oehl, and R. Höger, "The emotional spectrum in traffic situations: Results of two online-studies," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 18, pp. 168–188, 2013.

[20] G. Leu, N. J. Curtis, N. J. Curtis, and H. Abbass, "Modeling and simulation of road traffic behavior: artificial drivers with personality and emotions," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 17, no. 6, pp. 851–861, 2013.

[21] J. S. K. Ooi, S. A. Ahmad, Y. Z. Chong, S. H. M. Ali, G. Ai, and H. Wagatsuma, "Driver emotion recognition framework based on electrodermal activity measurements during simulated driving conditions," in *Proceedings of the 2016 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, pp. 365–369, Kuala Lumpur, Srilanka, December 2016.

[22] B. Hongzhe and S. Gongzhang, "Recognition and applications of emotion detection in driving fatigue," *The Open Automation and Control Systems Journal*, vol. 7, no. 1, 2015.

[23] L. Xie, Z. Wang, R. Dongchun, and S. Teng, "Research of driver emotion model under simplified traffic condition," *Acta Automatica Sinica*, vol. 36, no. 12, pp. 1732–1743, 2010.

[24] M. A. Neerincx, M. Harbers, D. Lim, and V. van der Tas, "Automatic feedback on cognitive load and emotional state of traffic controllers, engineering psychology and cognitive ergonomics," in *Proceedings of the International Conference on Engineering Psychology and Cognitive Ergonomics*, pp. 42–49, Berlin, Germany, June 2014.

[25] F. Riaz, S. Khadim, R. Rauf, M. Ahmad, S. Jabbar, and J. Chaudhry, "A validated fuzzy logic inspired driver distraction evaluation system for road safety using artificial human driver emotion," *Computer Networks*, vol. 143, pp. 62–73, 2018.

[26] J. Izquierdo-Reyes, R. A. Ramirez-Mendoza, M. R. Bustamante-Bello, S. Navarro-Tuch, and R. Avila-Vazquez, "Advanced driver monitoring for assistance system (ADMAS)," *International Journal on Interactive Design and Manufacturing*, vol. 12, no. 1, pp. 187–197, 2018.