IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 May 2024, accepted 30 June 2024, date of publication 3 July 2024, date of current version 3 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3422383

RESEARCH ARTICLE

A Multimodal Driver Anger Recognition Method Based on Context-Awareness

TONGQIANG DING[1], KEXIN ZHANG[1], SHUAI GAO[2], XINNING MIAO[3], AND JIANFENG XI[1]
[1]Transportation College, Jilin University, Changchun 130022, China
[2]Jilin Communications Polytechnic, Changchun 130022, China
[3]Beijing Jingwei Hirain Technologies Company Inc., Beijing 100029, China

Corresponding author: Shuai Gao (15698070@qq.com)

This work was supported in part by the National Key Research and Development Program Project "Major Accident Risk Prevention and Emergency Avoidance Technology for Road Transport Vehicles" under Grant 2023YFC3009600 and in part by the Graduate Innovation Fund of Jilin University.

ABSTRACT In today's society, the harm of driving anger to traffic safety is increasingly prominent. With the development of human-computer interaction and intelligent transportation systems, the application of biometric technology in driver emotion recognition has attracted widespread attention. This study proposes a context-aware multi-modal driver anger emotion recognition method (CA-MDER) to address the main issues encountered in multi-modal emotion recognition tasks. These include individual differences among drivers, variability in emotional expression across different driving scenarios, and the inability to capture driving behavior information that represents vehicle-to-vehicle interaction. The method employs Attention Mechanism-Depthwise Separable Convolutional Neural Networks (AM-DSCNN), an improved Support Vector Machines (SVM), and Random Forest (RF) models to perform multi-modal anger emotion recognition using facial, vocal, and driving state information. It also uses Context-Aware Reinforcement Learning (CA-RL) based adaptive weight distribution for multi-modal decision-level fusion. The results show that the proposed method performs well in emotion classification metrics, with an accuracy and F1 score of 91.68% and 90.37%, respectively, demonstrating robust multi-modal emotion recognition performance and powerful emotion recognition capabilities.

INDEX TERMS Context-awareness, driving state emotion recognition, emotional expression heterogeneity, multimodal emotion recognition, machine learning.

I. INTRODUCTION

In recent years, with the evolution of human-computer interaction and intelligent transportation systems, the application of biometric technology in the field of driver emotion recognition has garnered widespread attention. It assesses the emotional state of drivers by analyzing data such as facial expressions, speech information, physiological signals, and driving behavior, demonstrating significant application value in improving road safety and driving experience.

Emotions are experiences and subjective perceptions that humans generate in response to the external environment and events, along with corresponding behavioral reactions and changes. Different emotions can impact personal or social behavior in various ways. For drivers, emotional fluctuations directly affect the reception and judgment of road information during driving, thereby distracting the driver and affecting operational accuracy. Anger is one of the most common emotions experienced by drivers while driving. In China, angry driving is quite common, with data showing that about 68% of motor vehicle drivers have experienced "road rage." Parkinson [1] found that people are more likely to get angry while driving than not driving. Additionally, factors like time pressure and traffic congestion during driving increase the likelihood of anger while driving [2]. Angry driving is a serious public issue, with studies proving that anger during driving can have a range of adverse effects [3]. Cognitively, anger can negatively impact attention, perception, and information processing, affecting the driver's control over the vehicle. Behaviorally, driving

The associate editor coordinating the review of this manuscript and approving it for publication was Shaohua Wan.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

under the influence of anger can lead to more aggressive behavior [4].

In addressing the issues caused by angry driving, it is essential to start from the perspective of the individual driver and study the recognition of angry driving to facilitate early emotional soothing and safety warnings. This can prevent many traffic problems caused by angry driving, provide more precise protection for drivers and vehicles, and offer feasible safety assurances for the driving community. Therefore, this paper discusses a multi-modal method for recognizing driver anger emotions, with the remaining sections organized as follows. Chapter 2 introduces the current state of research on both unimodal and multimodal driver anger emotion recognition and elaborates on the problems this paper aims to solve. Chapter 3 focuses on the main innovations of the multi-modal anger emotion recognition method used in this paper. Chapter 4 introduces the dataset used in this study and related anger emotion recognition experiments. In terms of datasets, this paper emphasizes a multimodal dataset of facial expressions, voice, and vehicle driving states collected by the authors. Regarding recognition experiments, the paper not only validates the proposed method through experiments but also includes comparative and ablation experiments. Finally, Chapter 5 summarizes the experimental results and provides an outlook on future work.

II. RELATED WORK

In recent years, with the rapid development of sensor technology and machine learning, significant progress has been made in the recognition of driver anger emotions, both in unimodal methods, such as facial expression, speech information, physiological information, and behavioral information, and in multimodal methods involving the fusion of various types of information. Recognition methods based on facial expressions and speech information have become mainstream in emotion recognition due to their ease of data acquisition and high accuracy rates. In the field of driver anger emotion recognition, methods based on vehicle driving state information demonstrate unique data advantages. However, recognition methods based on physiological information, due to the challenges of non-contact data collection, often significantly impact the driver during the data collection process. Therefore, this paper will next elaborate in detail on the methods of anger emotion recognition based on facial information, speech information, and vehicle driving state information.

A. ANGER EMOTION RECOGNITION BASED ON FACIAL INFORMATION

Facial expressions are the most natural and direct way for humans to express emotions. Since facial information can be collected non-invasively, thus minimizing interference with the driver, facial emotion recognition has become the primary method for recognizing driver anger. The two most important steps in facial information-based anger emotion recognition are first to extract features from the facial information and then to determine whether the emotion is anger or not, using certain rules or classification models.

In terms of facial feature extraction methods, Ekman and Friesen [5] proposed the Facial Action Coding System (FACS) facial behavior coding method in 1978, which can be used to judge facial expressions by identifying specific Action Unit (AU) feature vectors. Barmana and Dutta [6] proposed a fast Sic active appearance model that can extract facial features by forming a mesh of 14 facial feature points. In addition, there are Active Shape Model (ASM), Local Binary Patterns (LBP) [7] and Histogram of Oriented Gradient (HOG) [8]. The purpose of the ASM algorithm is to generate points that fit the contours of the object. LBP, on the other hand, constructs a binary number by comparing each cell's pixels with its eight neighboring pixels through a function. Afterwards, the algorithm constructs a histogram of these numbers for each cell and connects the histograms. HOG is to produce a histogram based on the gradient previously calculated after segmenting the image into cells. However, in real life applications, all of the above methods have some drawbacks and shortcomings. The FACS system is more complex, and manually performing the FACS coding is a very time-consuming process, and the process is affected by the subjective judgment of the coder; Sic and LBP models are difficult to accurately capture the subtle dynamic changes in facial expressions; ASM's effectiveness relies heavily on the accuracy of the initial shape and is not flexible enough to deal with non-standard facial expressions; HOG is also difficult to accurately capture the subtle changes, and its computation may be relatively complex and time-consuming. In addition to the above methods, we observe that more feature extraction is performed by convolutional neural networks. Compared to the above methods, Convolutional Neural Networks (CNNs) can automatically learn and extract features from training data and realize hierarchical feature learning, which not only can extract higher-level features and subtle features that may be neglected by traditional methods, but also can improve the adaptability and flexibility of the model.

In terms of classification algorithms, there are two main categories: classical methods and neural network methods. Classical methods include methods such as support vector machines (SVMs), dynamic Bayesian networks, extreme learning machines [9], sparse learning machines [10], support vector regression [11], sparse representation classification [10], random forests [12], random trees [12], multi-graph embeddings [13], and single-modified Viola-Jines [14]. The above classical methods usually have deficiencies in terms of computational complexity, parameter tuning and model generalizability. DBN and random forests, for example, face the challenge of computational resources when dealing with large datasets, and the optimization of parameters such as the type of kernel function and regularization parameter in SVM, the number of nodes in the hidden layer in ELM, and the number of trees in random forests usually requires a lot of experiments and expertise, among other deficiencies. The main methods designed through neural networks include methods such as

multilayer perceptron (MLP) and CNN. Although MLP is a powerful classification tool, it still has some limitations in terms of deep structure design, overfitting control, parameter tuning, and dependence on data, etc. For example, MLP is unable to effectively utilize the spatial or temporal structure in the input data; MLP is susceptible to overfitting phenomenon when the training data is limited or noisy. CNN, on the other hand, can effectively capture the spatial hierarchical structure in the input data through its convolutional layer, and CNN usually has fewer parameters due to weight sharing and pooling operations, which reduces the computational complexity and the risk of overfitting, and improves the generalization ability of the model.AT Lopes [15] designed a multilayered CNN model that automatically extracts features and recognizes the emotion of anger from facial images; Liu et al. [16] used a three-dimensional CNN (3D-CNN) to capture spatio-temporal features in facial expression videos, not only focusing on the spatial features of the face, but also taking into account the change of expression over time, which improves the accuracy of emotion recognition; Ng et al. [17] proposed a CNN model that simultaneously learns facial emotion recognition and other face-related tasks (e.g., gender recognition, age estimation), showing that the potential that sharing representations between related tasks can improve emotion recognition performance. CNNs show great potential in facial emotion recognition, however, traditional CNNs usually contain a large number of parameters, leading to high computational complexity and significant storage requirements, which are not conducive to real-time applications; and when image data processing is performed, it is usually not processed to see which parts are more important for the final prediction, which may lead to the model to be disturbed by non-feature information.

B. ANGER EMOTION RECOGNITION BASED ON SPEECH INFORMATION

Language, as a unique means of human communication, can directly reflect human emotions through its vocal characteristics. In 1983, Bezooijen and others explored the correlation between vocal features and different emotions, suggesting that statistical parameters of speech features could be used for emotion classification. However, in the application field of driver anger emotion recognition, speech data often faces challenges such as poor data usability and missing data. Therefore, speech emotion recognition is usually used as a supplementary method to enhance the overall accuracy of driver anger detection. Speech emotion recognition also involves two critical steps: feature extraction and emotion classification.

In the aspect of feature extraction, speech features can be categorized into prosodic features, timbral features, and spectral features. Each category encompasses specific related characteristics, as shown in Table 1.

Prosodic features primarily focus on the rhythm, intensity, speech rate, and pitch of speech. For instance, John et al. [18] conducted experiments using prosodic features for emotion recognition, proving their effectiveness in speech emotion detection. Timbral features reflect the quality of speech signals, measuring the intelligibility, purity, and distinctiveness of speech. Research by Nussbaum et al. [19] has found that fundamental frequency, a timbral characteristic, plays a significant role in emotion recognition. Spectral features represent the characteristics of the signal in the frequency domain, where emotional fluctuations cause variations in the spectral distribution of speech. Lalitha et al. [20] have suggested the role of cepstral coefficients in spectral features for emotion classification in enhancing human-computer interaction performance. Furthermore, some scholars have combined the above-mentioned speech characteristics for emotion recognition. For example, Zhou et al. [21] proposed a method using a fusion of MFCC and prosodic features to identify speech emotions. Their experiments showed that this fusion increased the accuracy rate by nearly 20% compared to using a single feature, with the accuracy of using only MFCC being 62.3% higher than using only prosodic features.

TABLE 1. Speech features.

Feature type	Relevant features
Prosodic features	Fundamental Frequency Related Features、Time-Related Features、Energy-Related Features
Timbral features	Glottal Parameters、Formant Related Features、Frequency Jitter and Amplitude Jitter
Spectral features	Linear Predictive Cepstral Coefficients、Mel-Frequency Cepstral Coefficients

In the field of emotion classification, models mainly include SVM [22], [23], Artificial Neural Network (ANN) [20], Hidden Markov Model (HMM) [24], CNN [25], Decision Tree [26], Long Short-Term Memory (LSTM) [27], and Recurrent Neural Network (RNN) [28] methods. SVM is suitable for clear classification problems in high-dimensional spaces but only performs well on small to medium-sized datasets. CNN and LSTM demonstrate excellent performance but are computationally expensive. ANN has strong learning capabilities but is prone to overfitting and sensitive to parameter selection. HMM excels in processing time-series data but is limited in handling nonlinear features. RNN is apt for sequential data but struggles with long sequences. Decision Trees are easy to understand but prone to overfitting.

Many methods in the field of speech emotion recognition have achieved excellent recognition effects, but there are still some shortcomings in application scenarios like recognizing the anger emotion of drivers while driving. Firstly, to improve the comprehensive recognition accuracy of drivers' anger, the speech emotion recognition module should enhance recognition efficiency to achieve real-time overall recognition. However, neural network models like RNN, which have high recognition accuracy, lack real-time performance, while models like SVM, which have high recognition efficiency, need improved accuracy. Secondly, in the scenario of recognizing

drivers' anger while driving, issues such as missing speech information and poor information usability exist. Finally, the selection of features in speech emotion recognition often focuses on one or a few features, lacking comprehensiveness in feature selection.

C. ANGER EMOTION RECOGNITION BASED ON VEHICLE DRIVING STATE INFORMATION

Vehicle driving state information directly reflects the driver's behavior during driving. Studies have proven that this information can also reveal the driver's anger. For example, Lei Hu [29] designed a scale to measure drivers' anger expressions and conducted a survey using this scale. The results indicated that when driving in an angry state, there was an increase in lane-changing frequency, and operations such as accelerating, braking, and steering became more frequent and intense. Zhong et al. [30] and others collected data on drivers' behavior when they were angry, showing that driving speed, honking frequency, and incidences of speeding increased, while the behavior of slowing down at crosswalks decreased. Techer et al. [31] studied the impact of drivers' anger on attention processes and driving performance. The results suggested that anger affects driving behavior and attention, leading to significant fluctuations in driving behavior and decreased attention. Precht et al. [32] researched how drivers' anger impacts their driving behavior, analyzing data on drivers who showed anger towards driving errors, violations, and aggressive expressions, and comparing it with data from drivers who did not exhibit anger. The results indicated that anger led to more frequent aggressive driving behaviors, but did not increase the frequency of driving errors.

Based on these studies, some scholars have used vehicle driving state information for recognizing driver anger. For instance, Shafaei et al. [33] integrated the vehicle's yaw angle and acceleration into an emotion recognition system, creating two modules: a sudden car operation counter based on steering wheel rotation and an aggressive driving predictor based on acceleration changes. Combined with a facial emotion recognition module, the final result showed a 94% accuracy rate in predicting drivers' emotions. Wang [34] utilized vehicle motion information such as speed and steering wheel angle, along with electrocardiogram signals, for multimodal anger emotion recognition, achieving an accuracy rate of 84.75%. Yu [35] collected vehicle motion information like speed, acceleration, and steering wheel turning amplitude through a driving simulator, combining it with facial data for multimodal anger emotion recognition. Wang [36] considered more comprehensive vehicle driving state information through a driving simulator, including steering wheel angle, longitudinal and lateral speed, pitch angle, yaw angle, and engine speed, achieving an anger recognition accuracy rate of 65.8%.

Currently, research using driving state information for anger emotion recognition is not widespread and has some limitations. Firstly, most studies do not consider the impact of driver heterogeneity. Anomalies in vehicle driving state, such as significant changes in acceleration, may not necessarily indicate anger but could also be due to an aggressive driving style. Secondly, the anger reflected in the interaction between vehicles, such as vehicle following distance and frequency of overtaking, has not been captured.

D. MULTIMODAL ANGER EMOTION RECOGNITION BASED ON MULTIPLE INFORMATION FUSION

Due to the fact that unimodal methods can only provide one type of emotional information and in some situations, expressions of certain modalities may be suppressed [37], they have significant limitations in anger emotion recognition. Multimodal approaches, by considering a more comprehensive range of emotional expression channels, demonstrate better recognition performance. However, current multimodal research in the driver domain is relatively scarce, and there is no unified dataset for scholars to use. Therefore, both the selected modalities and the data used are diverse. For instance, Zhou et al. [38] proposed a multimodal fusion framework based on CNN+Bi-LSTM+HAM, which combines the driver's voice, facial images, and video sequences for emotion recognition, achieving an anger recognition accuracy of 85.0%. Ni et al. [39] collected physiological response signals, nasal tip temperature signals, and vehicle behavior signals from drivers in a simulated driving situation, conducting combination experiments with different data and methods using Random Forest (RF), K-Nearest Neighbor (KNN), XGBoost. The results showed that using the RF model for multimodal recognition of the three types of data was the most effective, with an anger recognition accuracy of 92.4%. Du et al. [40] proposed a Convolutional Bidirectional Long Short-Term Memory Neural Network (CBLNN), predicting the driver's emotions based on geometric features extracted from facial skin information and heart rate extracted from changes in RGB components, with an anger recognition accuracy of 90.5%.

In fact, multimodal fusion mainly falls into two categories: feature fusion and decision-level fusion. Decision-level fusion, due to its simplicity and tolerance to different modality recognition, often exhibits better performance. For example, Wang [36] conducted multimodal anger emotion recognition based on EEG signals, physiological signals, and driving behavior information. The experiments showed that decision-level fusion performed better than feature-level fusion, with an accuracy of 76.3%. Wang [34] conducted multimodal fusion recognition of anger emotions based on ECG signals and driving behavior signals. The results showed that the SVM-DS model, which employed decision-level fusion, performed best, with an accuracy of 84.75%. In previous studies, the weights of each modality in decision-level fusion were often fixed. However, in actual scenarios, the weights of modalities should vary. Li et al. [41] research found that drivers' emotional expressions are influenced by driving tasks, affecting emotion recognition. Tang et al. [37]

discovered that the distribution of multimodal physiological responses varies across different emotional scenarios. Specifically, the degree of expression in facial emotion, vocal emotion, and vehicle driving state emotion of drivers may differ under conditions such as congested versus smooth traffic flow, or when driving alone versus with passengers.

In summary, previous research work has made significant progress, but there are still some challenges and limitations. First, the recognition of driver emotions is a complex task requiring the consideration of multimodal data. However, issues such as individual differences among drivers and variability in emotional expression across different driving scenarios can impact the accuracy of emotion recognition. Second, while facial and vocal emotion recognition are the mainstream methods for emotion recognition, there is still room for improvement in their accuracy. Third, within our research scope, no current recognition methods consider using traffic flow information, which represents vehicle interaction, for anger emotion recognition, thus failing to fully capture the drivers' emotional expressions. Lastly, most scholars base their model training and validation on simulated scenarios, which still differ from real-world driving.

Based on this, this paper proposes a context-aware multimodal driver emotion recognition method (CA-MDER) aimed at overcoming these issues and effectively recognizing drivers' anger emotions. The contributions of this paper are as follows:

(1) A facial emotion recognition method based on AM-DSCNN is proposed, which introduces an attention mechanism module and a depthwise separable convolution module to enhance the model's ability to capture important facial emotional features and improve computational efficiency;

(2) A hybrid kernel function combining Dynamic Time Warping (DTW) with RBF is proposed to improve the SVM model, making the improved model more effective in handling the temporal elasticity in voice data and improving recognition accuracy;

(3) In the vehicle driving state emotion recognition module, features capturing vehicle interactions are introduced, and a driver driving style recognition module is proposed to mitigate the impact of driving heterogeneity on anger emotion recognition;

(4) A context-aware, multi-modal decision-level fusion method based on CA-RL is proposed, which uses context awareness to achieve optimal weight distribution in adaptive scenarios;

(5) This paper uses data collected from real driving scenarios and public datasets for model training and validation, which, compared to simulated data, results in a more realistic and effective model.

III. METHOD

For the recognition of driver anger emotions, a recognition framework that combines context-awareness and multimodal fusion is proposed, as shown in Figure 1. It is divided into the following parts:

(1) A facial anger emotion recognition module based on AM-DSCNN;

(2) A vocal anger emotion recognition module based on an improved SVM;

(3) A vehicle driving state emotion recognition module that considers vehicle driving state information and driving style;

(4) A multimodal decision-making module with adaptive weight distribution based on context awareness.

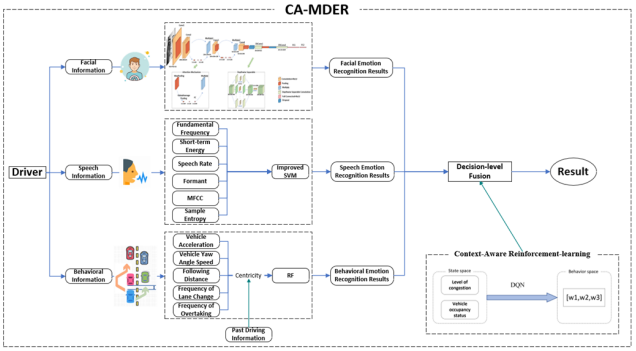


FIGURE 1. CA-MDER modeling framework.

A. FACIAL ANGER EMOTION RECOGNITION BASED ON AM-DSCNN

In response to the numerous shortcomings of traditional CNNs, this paper proposes a facial emotion recognition method based on Attention Mechanism-Depthwise Separable Convolutional Neural Networks (AM-DSCNN), specifically for recognizing the angry emotional states of drivers. The proposed AM-DSCNN is composed of three parts: the backbone network, depthwise separable convolution module, and attention mechanism module, the specific model structure is shown in Figure 2.

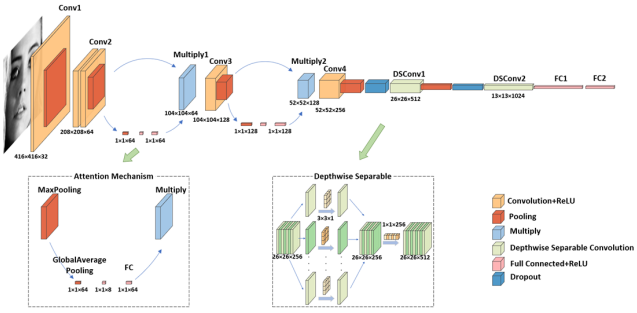


FIGURE 2. AM-DSCNN model structure diagram.

1) BACKBONE NETWORK

Backbone Network is the foundation of the AM-DSCNN model, consisting of convolutional layers, pooling layers, fully connected layers, and dropout layers. Its purpose is

to capture image features from low-level to high-level and ultimately provide emotion classification.

a: CONVOLUTIONAL LAYER

The convolutional layer is the core component of CNN, used to extract features from the input image. Convolution operations slide convolution kernels over the input image to perform dot product operations, resulting in feature maps. For example, in the first convolutional layer shown in Figure 2, the input image size in the model is 416 × 416×1. Then, the convolutional layer Conv1 with 32 convolution kernels of size 3 × 3 is used for convolution operations, and the output size after convolution is 416 × 416×32.

b: POOLING LAYER

The pooling layer is used to reduce the size of the feature maps, decrease computational complexity, and retain important features. For example, in the first pooling layer shown in Figure 2, the pooling layer uses a 2 × 2 window to perform max pooling on the feature map of size 416 × 416×32. The output size after pooling is 208 × 208×32.

c: FULLY CONNECTED LAYER

The fully connected layer connects all nodes from the previous layer to all nodes in the current layer, functioning as a standard neural network layer. As shown in Figure 2, this model includes fully connected layers in the attention mechanism module and at the final emotion classification. The former is used to generate attention weights for the image after global average pooling, thereby rescaling the feature maps; the latter is used to integrate features extracted from all previous layers for the final classification of angry emotions.

d: DROPOUT LAYER

The Dropout layer is a regularization technique used to prevent neural networks from overfitting. By randomly ''dropping out'' a portion of neurons during training, the network is forced to train on different sub-networks, thus enhancing the model's generalization ability.

2) DEPTHWISE SEPARABLE CONVOLUTION MODULE

When applying facial emotion recognition in a driver scenario, it's crucial to consider not only accuracy but also efficiency. For this reason, this paper introduces depthwise separable convolution into the model, a highly efficient convolution operation comprised of two steps: depthwise convolution and pointwise convolution.

a: DEPTHWISE CONVOLUTION

This involves applying a single filter to each input channel independently for convolution. Taking the first layer of depthwise separable convolution DSConv1 in Figure 2 as an example, a 3 × 3×1 convolution kernel is used to perform independent convolution on each channel of the input feature map with a size of 26 × 26×256. After the individual convolution operations, the resulting feature map has a size of 26 × 26×256.This process can be described as follows:

$$\boldsymbol{I}_{\boldsymbol{c}}^{'}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{I}_{\boldsymbol{c}}(\boldsymbol{x},\boldsymbol{y}) * \boldsymbol{K}_{\boldsymbol{c}} \tag{1}$$

Here, x, y are the spatial positions in the driver's facial feature map, I_c' is the output of the cth channel, I_c is the input of the cth channel, and K_c is the convolution kernel of the cth channel.

b: POINTWISE CONVOLUTION

This uses a 1 × 1 convolution kernel to perform convolution across channels, combining the outputs of the depth convolution. Similarly, for the first layer of depthwise separable convolution DSConv1 in Figure 2, after channel-wise convolution, a 1 × 1×256 convolution kernel is used to perform convolution, converting 256 channels into 512 channels. The final output feature map has a size of 26 × 26×512.This process can be described as follows:

$$\boldsymbol{I}_{\boldsymbol{c}}^{''}(\boldsymbol{x},\boldsymbol{y}) = \sum\nolimits_{\boldsymbol{c}} \boldsymbol{I}_{\boldsymbol{c}}^{'}(\boldsymbol{x},\boldsymbol{y}) * \boldsymbol{K}_{\boldsymbol{c}}^{'} \tag{2}$$

Here, K_c' is the cth channel of the 1 × 1 convolution kernel.

Compared to traditional convolution operations, depthwise convolution significantly reduces the number of multiplicative operations since each kernel only convolves on its respective single channel. Although pointwise convolution involves all channels, the computational load is still much lower than traditional convolution due to the use of 1 × 1 kernels. Furthermore, depthwise separable convolution requires far fewer parameters than classical convolution. For the same size of feature maps and convolution kernels, depthwise separable convolution can significantly reduce the model's parameter count, thereby lowering the risk of overfitting and enhancing the model's applicability in resource-constrained environments. Simultaneously, by decomposing standard convolution into these two steps, depthwise separable convolution still effectively captures the spatial patterns and texture information within each channel of the facial feature map and the feature combinations across channels, ensuring efficient feature extraction.

3) ATTENTION MECHANISM MODULE

Traditional CNNs struggle to identify which features are important for predictions, leading to the model being distracted by non-feature information. To address this issue, this paper introduces an attention mechanism module to help the model focus more on features crucial for emotion recognition. The attention mechanism module consists of a lightweight attention network, capable of adaptively adjusting the channel weights of the feature maps, directing the model's focus to features more beneficial for emotion recognition.

Specifically, this module is implemented through the following steps:

(1) Global Average Pooling: This paper applies global average pooling to the feature maps of each channel, reducing them to a scalar, essentially compressing to the channel dimension to obtain global context information, enabling

each channel feature to have a global perspective. For instance, in the first layer of the attention mechanism shown in Figure 2, global average pooling is performed on the input feature map with a size of 104 × 104×64 as per the following formula. After pooling, the output feature map size is 1 × 1×64.

$$C_{global} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} I(x, y) \tag{3}$$

Here, H and W are the height and width of the facial feature map.

(2) Multi-Layer Perceptron (MLP) Learning: The paper uses a multi-layer perceptron with one hidden layer to learn the non-linear dependencies between channels. This can be represented as:

$$C_{weight} = \delta(W_2 \delta(W_1 C_{global} + b_1) + bmb_2) \tag{4}$$

Here, W_1 and W_2 denote the weights in the MLP, b_1 and b_2 denote the biases in the MLP, and δ represents the ReLU activation function.

(3) Scaling Transformation and Feature Map Re-Calibration: The sigmoid activation function is applied to obtain the channel attention weights, and the original feature map is re-calibrated accordingly:

$$Z = \sigma(C_{weight}) \odot I \tag{5}$$

Here, σ is the sigmoid activation function, and Z is the feature map after attention re-calibration.

B. SPEECH ANGER EMOTION RECOGNITION BASED ON IMPROVED SVM

1) EXTRACTION OF EMOTIONAL FEATURES FROM SPEECH INFORMATION

Effective extraction of speech features is crucial for identifying drivers' angry emotional states based on speech information. To comprehensively extract emotional characteristics, this paper selects six key features: fundamental frequency, short-term energy, speech rate, formants, Mel-Frequency Cepstral Coefficients (MFCC), and sample entropy. The features extracted from preprocessed speech signals are frame-level and only represent local emotional characteristics of the speech signal. To extract global features, it is necessary to calculate the global statistical properties of multi-frame speech signals to obtain utterance-level features. The statistical measures derived from sample data can reflect the quantitative characteristics of the sample population. For different speech features, their respective parameter statistics are calculated for subsequent emotion recognition.

a: FUNDAMENTAL FREQUENCY FEATURES

The fundamental frequency, related to the vocal cord status, varies under different emotions, thus making it a good representation for speech emotion. The formula for calculating the fundamental frequency using the autocorrelation method is:

$$F0 = argmax\,(R(k)) \tag{6}$$

$$R\,(k) = \sum_{n=1}^{N-k-1} x\,(n) \cdot x\,(n+k) \tag{7}$$

Here, F0 is the fundamental frequency, $R\,(k)$ is the autocorrelation function, $x\,(n)$ is the signal value at time n, N is the frame length, and k is the delay amount.

b: SHORT-TERM ENERGY FEATURES

Literature [42] confirms the effectiveness of short-term energy E in speech emotion recognition, which is calculated by summing the squares of the sample points within the frame:

$$E = \sum_{n=0}^{N-1} x(n)^2 \tag{8}$$

c: SPEECH RATE

Speech rate refers to the number of speech units spoken per unit of time and is one of the important features of speech signals. Speech rate not only reflects the speaker's linguistic style and habits, but is also closely related to his/her emotional state. Literatures [43] and [44] suggests that high arousal emotions such as anger and excitement are usually accompanied by faster speech rate. Specifically, anger tends to cause speakers to speak faster because of the increased physiological arousal level in anger, which leads to rapid breathing and faster speech tempo.

d: FORMANT FEATURES

The quality of speech, affected by vocal tract deformation under different emotional states, exhibits distinct feature changes. Therefore, the peak values and positions of formants in speech signals vary under different emotional states. This paper uses Linear Predictive Coding (LPC) to extract the central frequencies of the first three formants, followed by peak detection estimation on the LPC spectrum.

e: MFCC

The Mel Frequency Cepstrum Coefficient (MFCC) combines the auditory perceptual properties of the human ear with the generation mechanism of speech signals, and can be converted between frequencies (in Hz) and Mel frequencies. The conversion formula is as follows, where f represents the frequency of the speech signal at 16,000 Hz.

$$Mel\,(f) = 1125ln\left(1 + \frac{f}{700}\right) \tag{9}$$

It is shown that the data of the speech signal is mainly concentrated in the low frequency region after the transform, so it is sufficient to extract the first 12 MFCC coefficients as features. After the frame-splitting operation, the global sentiment features need to be extracted. The static features of MFCC are logarithmic energy, while the dynamic features can be obtained by calculating the derivatives of the static features. Therefore, in this paper, instead of directly extracting the mean value of the logarithmic energy, the first-order

derivative and second-order derivative of the mean value of the logarithmic energy of the MFCC are calculated. The formula for the first-order derivatives is as follows, and the second-order derivatives are calculated similarly to the first-order derivatives, where s is the range to find the difference and is taken as $s = 2$.

$$\Delta C_n(m) = \frac{\delta}{\delta m} C_n(m) \cong \frac{1}{T_S}\left[\sum_{t=-s}^{s} t C_{n-1}(m)\right] \tag{10}$$

$$T_S = \sum_{t=-s}^{s} t^2 \tag{11}$$

By combining the dynamic and static features of MFCC, the performance of the speech emotion recognition model can be effectively improved.

f: SAMPLE ENTROPY

Sample entropy (SampEn) is a statistical measure for quantifying the complexity of time series, proposed by Richman and others in 2000. It is particularly effective in analyzing nonlinear dynamic systems, such as speech signals, because it quantifies the complexity of time series from a probabilistic perspective. Compared to approximate entropy, sample entropy does not include self-comparison in its calculations, reducing data bias. Even with limited data, sample entropy can effectively estimate probabilities, hence providing higher detection accuracy [45]. Sample entropy is defined as the natural logarithm of the conditional probability that data vectors of dimension m remain similar when the dimension increases to $m + 1$. Specifically, sample entropy SampEn(m,r,N) is defined as:

$$SampEn(m, r, N) = -\ln \frac{B^{m+1}(r)}{B^{m}(r)} \tag{12}$$

$$B^{m}(r) = \frac{1}{N - m + 1} \sum_{i=1}^{N-m+1} B_i^{m}(r) \tag{13}$$

Here, m is the embedding dimension, r is the similarity threshold for judging whether two sequences are similar, and $B^{m+1}(r)$ and $B^{m}(r)$ are the normalized counts of similar sequence pairs at different dimensions. $B_i^{m}(r)$ represents the similarity of a specific speech data vector with other vectors in the m-dimensional space. The calculation process can be described as: for a given embedding dimension m and similarity threshold r, calculate the similarity of a specific speech data vector $X(i)$ with all other vectors $X(j)$ in the m-dimensional space and count the number of similar vector pairs.

In the application of speech emotion recognition, sample entropy can reveal the complexity and dynamic changes of speech signals under different emotional states. For example, angry or excited speech may exhibit higher sample entropy, indicating more complex and variable signals, while calm or sad speech may have lower sample entropy, reflecting more stable and consistent characteristics. Therefore, incorporating sample entropy as a feature into the emotion recognition model can help us more accurately capture and differentiate speech features under different emotional states, thereby enhancing the accuracy and efficiency of emotion recognition.

2) SPEECH EMOTION RECOGNITION MODEL BASED ON IMPROVED SVM

SVM has been proven in the literatures [46] and [47] to have better recognition performance in speech emotion recognition (especially for angry emotions), which, together with its significant advantages in computational efficiency and memory usage, makes it suitable for use as a as a complementary recognition for facial emotion recognition in multimodal emotion recognition tasks. In this paper, based on the traditional SVM model, considering the time series characteristics of speech data, we improve the SVM kernel function and propose a hybrid kernel function combining dynamic time warping (DTW) and Radial Basis Function (RBF). The DTW algorithm can effectively deal with temporal elasticity on time-series data because it is able to bend the time axis in order to find the best correspondence between two time-series, which means that even if the speech samples are different in time, it can find the similarity between them, thus improving the accuracy of the model for sentiment classification.

The improved hybrid kernel function $K(x, y)$ is defined as:

$$\begin{aligned} K(x, y) &= \alpha \cdot K_{DTW}\left(x_{i,t}, x_{j,t'}\right) + (1-\alpha) \cdot K_{RBF}\left(x_i, x_j\right) \\ &= \alpha \cdot \exp(-\gamma_{DTW} \cdot DTW(x_{i,t}, x_{j,t'})) \\ &\quad + (1-\alpha) \cdot \exp(-\gamma_{RBF} \cdot \|x_i - x_j\|^2) \end{aligned} \tag{14}$$

where K_{DTW} is the DTW-based kernel, used for time-series features, with i and j representing the feature and label data, respectively, and $x_{i,t}, x_{j,t'}$ is the value of the time series at a certain point in time (such as fundamental frequency, short-term energy, speech rate, and formant features); K_{RBF} is the Radial Basis Function (RBF) kernel, used for statistical features, where x_i, x_j are two data points in the feature space (such as MFCC, sample entropy features);α is a parameter to adjust the weight of the two, and γ_{DTW} and γ_{RBF} are the respective scaling parameters of the kernels.

The computation formula for DTW can be expressed as:

$$DTW(A, B) = min\left(\sqrt{\sum_{t=1}^{T} (x_{i,t} - x_{j,t'})^2}\right) \tag{15}$$

C. DRIVING STATE EMOTION RECOGNITION BASED ON RF

1) EXTRACTION OF EMOTIONAL FEATURES FROM VEHICLE DRIVING STATE INFORMATION

In the recognition of anger emotion based on vehicle driving state, this paper not only selects vehicle acceleration and yaw rate, which represent the vehicle's motion parameters, but also for the first time proposes the use of three behavioral features representing vehicle-to-vehicle interaction: following distance, lane-changing frequency, and overtaking frequency.

These features are used to comprehensively extract the driver's driving state emotions for more accurate recognition of the driver's anger emotion.

a: VEHICLE ACCELERATION

The degree of speed change per unit time is known as acceleration, and its magnitude or rate of change directly affects the urgency of speed variation. Acceleration is a key indicator of the extent of vehicle speed change and reflects the driver's longitudinal control ability over the vehicle. Literature [36] indicates that as the intensity of the driver's anger increases, the fluctuation amplitude of vehicle acceleration also increases, leading to reduced smoothness in vehicle motion, i.e., the driver's longitudinal control ability over the vehicle decreases. Therefore, this paper extracts the mean value of vehicle acceleration a_{av} as one of the indicators for emotion recognition.

b: YAW RATE

When the driver is in a normal driving state, the range of changes in the vehicle's yaw rate is small, and the frequency of change is high. Under the state of anger, the range of acceleration change is larger, and the adjustment frequency is lower. Literature [36] shows that as the driver's anger intensity increases, the fluctuation amplitude of the vehicle's yaw rate also increases, leading to reduced lateral stability of the vehicle, i.e., the driver's lateral control ability over the vehicle decreases. Hence, this paper extracts the mean value of the vehicle's yaw rate $\dot{\psi}_{av}$ as one of the indicators for emotion recognition.

c: FOLLOWING DISTANCE

Following distance refers to the distance between vehicles traveling on the road. Generally, a reduced following distance allows vehicles to pass through intersections and other delay-prone areas more quickly, but as the following distance decreases, safety risks increase. Literature [48] indicates that the following distance tends to decrease under the driver's anger emotion. Therefore, this paper extracts the mean following distance g_{av} as one of the indicators for emotion recognition.

d: LANE-CHANGING FREQUENCY

Lane changing refers to a vehicle moving from its current lane to an adjacent lane on the road, often to avoid obstructions, pursue other vehicles, or turn at intersections. Lane changing can help the vehicle bypass obstacles or slow-moving vehicles in its current lane, maintaining a smooth speed. If one lane is faster than another, changing lanes can help the vehicle increase its speed and reach its destination more quickly. However, frequent lane changes can disrupt traffic flow and increase the risk of traffic accidents. Literature [29] indicates that the frequency of lane changes increases under the driver's anger emotion. Therefore, this paper extracts the frequency of lane changes cl_{fr} within a unit time window as one of the indicators for emotion recognition.

e: OVERTAKING FREQUENCY

Overtaking refers to a vehicle passing slower-moving vehicles ahead on the road, usually by moving into an adjacent lane. When a vehicle is moving fast, overtaking can help it quickly bypass the slower vehicles ahead, reducing traffic congestion. Reasonable overtaking can make traffic flow smoother and allow vehicles to travel at appropriate speeds, avoiding the formation of a slow-moving convoy. However, unsafe overtaking can lead to traffic accidents, especially in conditions of poor visibility, complex road conditions, or inappropriate timing for overtaking. On busy roads, frequent overtaking can disrupt traffic flow. Literature [49] shows that the overtaking frequency increases under the driver's anger emotion. Therefore, this paper extracts the frequency of overtaking otk_{fr} within a unit time window as one of the indicators for emotion recognition.

2) STATISTICAL ANALYSIS OF DRIVING STYLE BASED ON HISTORICAL DRIVING DATA

Due to individual differences in driving style (habits) or driving experience, driving behaviors may vary among different drivers. Therefore, in addition to the emotional (anger) state affecting the driving behavior characteristics of the driver, the individual differences between drivers can also have a certain impact on these characteristics, thereby affecting the accuracy of the anger recognition model for driving. However, the anger characteristics of the same subject have strong stability over different periods. Hence, this paper adopts a driving style analysis method based on historical driving data to reduce the impact of individual differences among drivers on the recognition of their anger state, thereby enhancing the robustness and accuracy of the recognition model.

For the feature subset obtained earlier $M = \{a_{av}, \dot{\psi}_{av}, g_{av}, cl_{fr}, otk_{fr}\}$, calculate the mean value of each feature under the normal state for each subject. This mean value is taken as the reference value. Let the mean value of the feature parameters of the ith subject under normal driving state be R_i, that is, the reference value. The equation is as follows:

$$\boldsymbol{R_i} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{M_i} \tag{16}$$

Finally, the feature parameter values of all subjects under normal state are inputted into the training sample library as the statistical values of each subject's driving style.

3) DRIVING STATE EMOTION RECOGNITION MODEL BASED ON RANDOM FOREST

The random forest model builds multiple decision trees and derives the final classification result by majority voting among these trees. This makes the random forest model

generally robust against outliers and noise and reduces the risk of overfitting. Moreover, random forests adapt better to unbalanced datasets. Based on these advantages, this paper considers the random forest model to be more suitable for the task of vehicle driving state emotion recognition. The implementation steps of the model are as follows:

(1) Feature Centralization

To better capture the emotional differences of drivers and reduce the impact of individual differences on anger state recognition, feature data must be centralized before applying the random forest model. This is done by obtaining the average feature value for each individual under normal conditions from the sample library, and then subtracting the actual feature value to obtain the centralized feature value C_i. The process can be represented as:

$$\boldsymbol{C_i = M_i - R_i} \tag{17}$$

where M_i is the actual value of the feature of the ith driver, and R_i is its average feature value.

(2) Bootstrap Sampling

Randomly select samples from the original dataset to create a new dataset. Let the size of the original dataset be N. Then N samples are selected from the original dataset with putback to form a new dataset D_i:

$$\boldsymbol{D_i = \left\{(x_1^*, y_1^*), (x_2^*, y_2^*), \cdots, (x_N^*, y_N^*)\right\}} \tag{18}$$

where D_i represents the training data for the ith tree, and (x_i^*, y_i^*) is the sample randomly selected from the original dataset.

(3) Decision Tree Construction with Feature Subsets

In terms of using driving state information for emotion recognition, the literature [39] has demonstrated that random forest models (RF) have good performance. Each decision tree is trained on its corresponding bootstrap sampling dataset D_i. However, in the splitting process of each node, not all features are considered, but a random feature subset is selected. Assuming the original number of features is M, m features are randomly selected at each node split.

(4) Decision Tree Ensemble

Using the above method, T decision trees are constructed in the random forest, each independently trained based on different bootstrap sampling datasets.

(5) Prediction

The prediction of the random forest is based on the predictions of all its decision trees. Specifically, for classification tasks, each tree provides a classification prediction, and the final prediction of the random forest is the mode of these classifications:

$$\hat{y} = \boldsymbol{mode}(\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_T) \tag{19}$$

Here, $\hat{y}_i$ represents the prediction of the ith tree, and mode represents taking the mode of these classification results, i.e., the most frequently occurring classification prediction.

D. ADAPTIVE MULTIMODAL ANGER EMOTION RECOGNITION BASED ON CA-RL

1) ADAPTIVE WEIGHT ALLOCATION BASED ON CA-RL

In multimodal emotion recognition, how to assign appropriate weights to each modality is the key issue. In order to make the system more adaptable to different driving environments and individual differences, we propose a method based on reinforcement learning to perceive the scenarios, find the optimal strategy in different scenarios by interacting with the environment, and dynamically assign weights to each modality, so as to improve the model's adaptability and generalisation ability.

a: MODEL DEFINITION

State space s_t: Emotion recognition scene, characterized by the degree of traffic congestion C_{tr} and the state of the number of people inside the vehicle H_{pe}. This includes the driver's emotions, vehicle state, traffic conditions, etc.

Their representation is as follows:

$$\boldsymbol{C_{tr} = \frac{N}{L}} \tag{20}$$

$$\boldsymbol{H_{pe} = \begin{cases} 1, & if\ \Delta E_t > e \\ 0, & otherwise \end{cases}} \tag{21}$$

Here, N is the number of vehicles at a given moment within the observed road section, L is the length of the observed section, ΔE_t is the short-term energy difference in voice between adjacent windows, and e is a threshold. The number of people inside the vehicle is detected by voice energy fluctuation. When a single person speaks, the short-term energy usually shows a more consistent pattern, whereas in multi-person conversations, energy fluctuations may be more pronounced and frequent. This is because different speakers usually have different voice intensities and rhythms, causing more peaks and valleys in energy levels during multi-person conversations.

Action space a_t: The weight parameters of the three modalities in the decision-level fusion of the anger recognition models, represented as $[w_1, w_2, w_3]$.

Reward function r_t: Given based on the consistency of the emotion recognition result with the actual emotion, defined as:

$$\boldsymbol{r_t = \begin{cases} 1, & if\ prediction\ matches\ actual\ emotion \\ 0, & otherwise \end{cases}} \tag{22}$$

b: REINFORCEMENT LEARNING ALGORITHM

This paper chooses the Deep Q Network (DQN) to implement dynamic weight adjustment. DQN attempts to estimate an action-value function $Q(s,a)$ representing the expected return of choosing action a in state s.

The core update formula of DQN is:

$$\boldsymbol{Q(s_t, a_t) = r_t + \gamma max_{a'} Q(s_{t+1}, a')} \tag{23}$$

Here, γ is the discount factor for future rewards, and $\max_{a'} Q(s_{t+1}, a')$ is the maximum estimate of the expected return for all possible actions a′ in the future state s_{t+1}.

2) DECISION-LEVEL FUSION BASED ON ADAPTIVE WEIGHTS

The emotion recognition results obtained by each modality for the emotion recognition sample x are $\{p_{le}(x), l = 1, 2, 3; e = 1, 2\}$, where l is the modality category, and e is the emotion category. The anger emotion recognition results of each modality are fused at the decision level, and the final probability expression after fusion is:

$$p'_e(x) = \frac{R\left(p_{le}(x)\right)}{\sum_{e=1}^{2} R\left(p_{le}(x)\right)} \tag{24}$$

Here, $p'_e(x)$ is the final probability of the sample data being recognized as emotion category e, and R is the fusion criterion, i.e., linear weighted fusion using the adaptive weights $[w_1, w_2, w_3]$ obtained previously. Therefore, the final expression for determining emotion classification is:

$$f(x) = \arg\max_e \left(p'_e(x)\right) \tag{25}$$

IV. DATA AND EXPERIMENT

A. DATA

The method proposed in this paper involves multimodal anger emotion recognition using AM-DSCNN, improved SVM, and RF models for facial, voice, and driving state information, respectively. Additionally, it employs an adaptive weight allocation based on CA-RL for multimodal decision-level fusion. For this recognition method, two types of datasets were selected. On one hand, to enhance the generalizability of the facial and voice recognition models, large public datasets such as AffectNet and SEWA were used. On the other hand, to utilize specific driving state features and the adaptive weight allocation method based on CA-RL, a multimodal dataset, Multimodal Data, was constructed by the team for training and validation of the model.

1) PUBLIC DATASETS

a: AffectNet

The AffectNet database [50], also known as a large-scale dataset for facial emotion recognition and analysis, was released in 2017. AffectNet contains over one million facial images collected from the internet using 1250 emotion-related keywords in six different languages and queries to three major search engines. It includes a variety of races, ages, and cultural backgrounds. Different emotions were filtered from the dataset to obtain data that meet the needs of this work, with the composition of emotion labels detailed in Table 2.

TABLE 2. Overview of the AffectNet dataset.

Emotion Category	Total Number of Samples	Number of Samples in Training Set	Number of Samples in Validation Set
Neutral	28858	23086	5772
Angry	19240	15392	3848
Total	48098	38478	9620

b: SEWA

The SEWA speech database [51] was released in 2019. It recorded 6 groups of volunteers (each group with 30 people) from six different cultural backgrounds: UK, Germany, Hungary, Greece, Serbia, and China. The gender and age distribution of each group of volunteers was wide-ranging. The database thus generated includes 199 experimental records, comprising 1525 minutes of audiovisual data recording the reactions of 398 individuals to advertisements, and over 550 minutes of computer-mediated interactions between subjects face-to-face. Different emotions were filtered from the dataset to obtain data that meet the needs of this work, with the composition of emotion labels detailed in Table 3.

TABLE 3. Overview of the SEWA dataset.

Emotion Category	Total Number of Samples	Number of Samples in Training Set	Number of Samples in Validation Set
Neutral	2000	1600	400
Angry	1264	1011	253
Total	3264	2611	653

2) MULTIMODAL DATASETS

Due to the current lack of vehicle driving state data under real vehicle operating scenarios, this paper designed a real vehicle experiment to collect multimodal data of drivers under different scenarios. The real vehicle operation location was set in the Jingyue District of Changchun City, Jilin Province, China. The operational scenarios were controlled by two intersecting factors: traffic flow density (with three scenarios: severely congested, generally congested, and uncongested) and the number of drivers inside the vehicle (either single or multiple). The experiment recruited 8 drivers, including 4 males and 4 females, with male ages ranging from 22-47 and female ages from 23-49.

a: DATA COLLECTION EQUIPMENT

To collect multimodal emotion information, this study used cameras and recorders to collect facial and speech information of drivers; drones were used to collect driving state information, especially the micro-traffic flow information representing vehicle interactions. The information collection equipment is detailed in Table 4.

b: EMOTION INDUCTION

This paper uses the emotion induction method from Literature [36], employing emotion induction materials relevant to the cultural background, for inducing anger and neutral emotions. The related materials are listed in Table 5. Drivers

TABLE 4. Data collection equipment.

Name	Quantities	Unit
FAW Volkswagen Magotan Sedan	1	Set
Drone (DJI Mavic Air2)	1	Set
Camera(Sony HDR-CX405)	1	Set
Recorder(Newsmy W6)	1	Set

watched related video materials before real vehicle operation, followed by 20 minutes of vehicle operation and data collection as one data collection cycle.

TABLE 5. Emotion induction materials.

Induced Emotion	Stimulus Duration	Induction Success Rate	Video Name
Anger	100s	86%	"Japanese Government Publicly Releases 'Unit 731' Roster for the First Time, Confirming the Existence of the 'Devil's Brigade'"
Neutral	96s	92%	"Domestic's Strongest Sedan SUV: 2019 Geely-Star 2.0T High Configuration!" Experimental Parameter Settings

c: DATA PREVIEW

The collected multimodal data were filtered, including the exclusion of driving state video data where the main vehicle was not captured and the removal of face video data where faces were obscured. This resulted in the Multimodal Data dataset, which includes 600 samples. Each sample consists of simultaneously recorded facial data, voice data, and vehicle driving state data, with the composition of emotion labels detailed in Table 6.

TABLE 6. Overview of the multimodal data dataset.

Emotion Category	Total Number of Samples	Number of Samples in Training Set	Number of Samples in Validation Set
Neutral	364	291	73
Angry	236	189	47
Total	600	480	120

B. EXPERIMENTAL SETUP

The experiments in this paper were conducted using Python 3.10.9 and the deep learning framework TensorFlow 2.14.0. In terms of the experimental environment, we used the Windows 10 operating system. For the experimental hardware, we used an Intel Core i7-13620H processor and a GeForce GTX 4060 graphics card. Parameter setting is particularly important in the model training process. Table 7 lists the parameter settings for our experiments.

TABLE 7. Experimental parameter settings.

Model	Parameter	Value
AM-DSCNN	Optimizer	Adam
	Learning rate	1e-4
	Dropout	0.15
	Batchsize	64
	Epoch	100
SVM	gamma	auto
	C	100
	tol	1e-3
RF	n_estimators	100
	max_features	auto
	min_samples_split	2
	min_samples_leaf	1
RL	Learning rate	1e-3
	Discount Factor	0.95
	Exploration Rate	0.9(Step down)
	Episodes	1000
	Step Size	200

C. EVALUATION METRICS

This study uses Accuracy and F1 score as the evaluation criteria for the model.

a: ACCURACY

Accuracy is the most intuitive performance metric, representing the proportion of correctly predicted instances to the total number of predictions. It reflects the credibility of the model's predictions. Its calculation formula can be expressed as:

$$\mathbf{ACC} = \frac{\mathbf{TP} + \mathbf{TN}}{\mathbf{TP} + \mathbf{TN} + \mathbf{FP} + \mathbf{FN}} \tag{26}$$

b: F1 SCORE

The F1 score is the harmonic mean of Precision and Recall. Its calculation formula is:

$$\mathbf{F1} = \frac{\mathbf{2} \times \mathbf{precision} \times \mathbf{recall}}{\mathbf{precision} + \mathbf{recall}} \tag{27}$$

D. SINGLE MODALITY AND CONTEXT-AWARE MULTIMODAL FUSION ANGER EMOTION RECOGNITION EXPERIMENT

1) FACIAL ANGER EMOTION RECOGNITION EXPERIMENT

Firstly, the dataset images were subjected to preprocessing operations such as normalization, grayscale processing, and image enhancement. The effect after preprocessing is shown in Figure 3 (b). Then, the AM-DSCNN facial anger emotion recognition model proposed in this paper was trained according to the parameters set in Table 7. After training, Grad-CAM was used to visualize the attention mechanism in the model structure as a heatmap. This was done by calculating gradients on the last convolutional layer and then rendering it as a heatmap overlaid on the original image, as shown in Figure 3 (c).

(a) Original figure (b) Pre-processed figure (c) Attention mechanism effect figure

FIGURE 3. Experimental correlogram for facial anger recognition.

Validation tests were conducted on the AffectNet validation set and the Multimodal Data validation set, with results for various indicators shown in Table 8.

TABLE 8. Performance results of the AM-DSCNN model on the dataset.

Dataset	ACC(%)	F1(%)
AffectNet	87.91	87.29
Multimodal Data	85.64	84.89

2) VOICE ANGER EMOTION RECOGNITION EXPERIMENT

The voice signals collected by the equipment are often affected by the environment, vehicles, and people, leading to many silent segments, strong and unstable noise. Therefore, to obtain a more uniform and smooth signal and extract more complete voice feature parameters, preprocessing of voice signals is necessary. This includes effective endpoint detection, pre-emphasis, framing, and windowing.

Endpoint detection refers to detecting the start and end points of a speech signal, which is essentially a two-classification problem of distinguishing the speech segment and the silence segment of a sample. The endpoint detection can reduce the influence of environmental noise, reduce the amount of system computation and computation time, and improve the system real-time. In this paper, we use the double threshold method based on short-time energy and short-time over-zero rate to achieve the endpoint detection of speech signals, and the idea of the double threshold method is shown in Figure 4. Firstly, the short-time energy is used for the first level of discrimination. Find the short-time energy E_k and the threshold value T_1 and T_2 for each frame; higher than T_2 is defined as the voice segment, the voice start and end point should be located outside the CD segment; therefore, from point C to the left, point D to the right, search for points B and E where E_k and T_1 intersect, and the BE segment is the start and end point of the voice segment of the first level of determination. Then use the short time zero rate for the second level of judgement. Find the short time zero rate Z_k and the threshold value T_3 for each frame; then search from point B to the left and point E to the right to find the points A and F where Z is lower than T_3, and the AF segment is the starting and stopping point of the speech segment of the second level of determination, i.e., the effective endpoint of the speech.

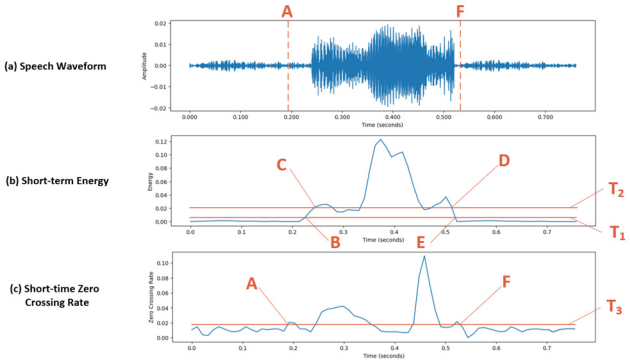


FIGURE 4. Effective voice endpoint detection diagram.

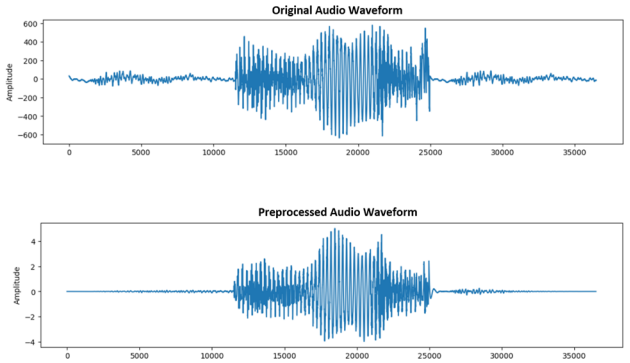


FIGURE 5. Voice data preprocessing effect diagram.

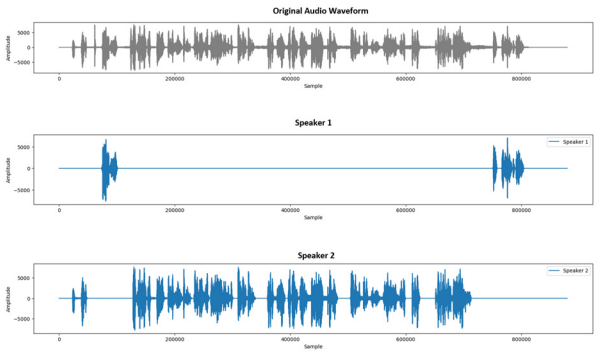


FIGURE 6. Speaker diarization clustering diagram.

After effective endpoint detection, voice data were preprocessed with pre-emphasis, framing, and windowing. The effect before and after preprocessing is shown in Figure 5.

In addition to this, in some scenarios will be collected to contain two or more people and mixed speech, so part of the voice data also needed to enter the driver's speech information directional extraction. In this paper, the open source tool falcon package is used to perform voiceprint segmentation and clustering (the effect graph after voiceprint segmentation and clustering is shown in Figure 6), and then the driver's voice data is selected, so as to complete the directional extraction of the driver's voice.

When speech features are extracted and computed, different combinations of parameters may have different effects on the computation of the features, and here we take short-time energy as an example to show the parameter sensitivity analysis in the experiment. In this paper, the parameter sensitivity experiment is carried out with the main parameters of short-time energy, such as window length and overlap rate, as shown in Figure 7.

In Figure 7, F stands for frame size and O stands for overlap size. From the results of the average energy and variance, the change of overlap size has a small effect on both the average energy and variance, which means that the short-term energy characteristics are not sensitive to the selection of the parameter of overlap size. On the other hand, the increase of frame size will not only increase the average energy, but also make the energy variance larger, which means that the short-term energy is more sensitive to the selection of the parameters of frame size.

Based on the preprocessing of speech data and parameter sensitivity analysis, the improved SVM voice anger emotion recognition model proposed in this paper was trained and validated. Tests were conducted on the SWEA training and validation sets, and the Multimodal Data validation set, with performance indicators involved shown in Table 9.

TABLE 9. Validation results of the improved SVM model.

Dataset	ACC(%)	F1(%)
SWEA	70.52	70.13
Multimodal Data	76.38	75.72

3) DRIVING STATE ANGER EMOTION RECOGNITION EXPERIMENT

In this study, drones were used to collect vehicle driving state information under different scenarios, flying at a height of 250 m during the collection process. In addition to this, vehicle driving status information can be obtained through on-board sensors and devices or road test data acquisition facilities. The Yolov5x algorithm was used for vehicle recognition and detection. First, the main vehicle was framed, followed by tracking and recognition using the detection algorithm. Then, combining the video frame rate with the real-time flight speed from the drone's log file, driving features were calculated, resulting in $M = \{a_{av}, \dot{\psi}_{av}, g_{av}, cl_{fr}, otk_{fr}\}$. The program operation is shown in Figures 8 and 9.

After calculating the various feature values of the vehicle's driving state, the Random Forest model was used to train and validate the Multimodal Data dataset. The final experimental results for each indicator are shown in Table 10.

TABLE 10. Performance results of the random forest model on the dataset.

Dataset	ACC(%)	F1(%)
Multimodal Data	68.74	65.53

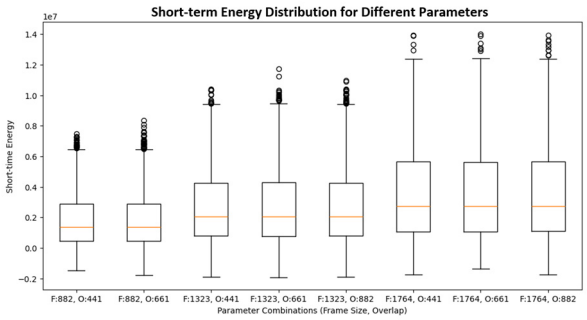


FIGURE 7. Voice data preprocessing effect diagram.

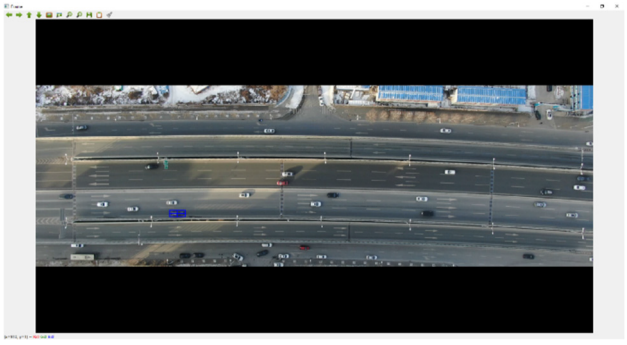


FIGURE 8. Main vehicle framing schematic diagram.

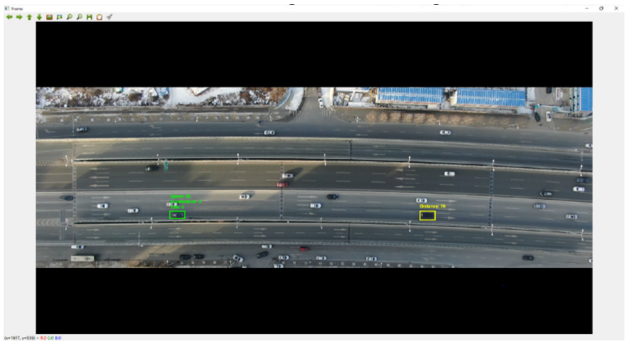


FIGURE 9. Tracking identification and calculation of driving status information.

E. CONTEXT-AWARE MULTIMODAL FUSION ANGER EMOTION RECOGNITION COMPARATIVE EXPERIMENT

This study determined the real-time driving scenario through traffic density detection and voice energy fluctuation detection, defining the state space accordingly. The conceptual diagram of perception is shown in Figure 10. For traffic density detection, the Yolov5x algorithm was used to detect continuous lane lines closest to the main vehicle. After successful detection, vehicles within the lane lines were identified to obtain traffic density information. For voice energy fluctuation detection, analysis was conducted based on the short-term energy difference between adjacent windows in the voice features.

FIGURE 10. Traffic density perception schematic diagram.

Next, reinforcement learning was used to train and validate different scenarios in the action space (i.e., weight schemes) on the Multimodal Data dataset. The performance indicators of the CA-MDER model proposed in this paper for the multimodal emotion recognition task of drivers are shown in Table 11.

TABLE 11. Performance results of the CA-MDER model on the dataset.

Dataset	ACC(%)	F1(%)
Multimodal Data	91.68	90.37

Due to the diversity of modalities and data types used in multimodal emotion recognition, there is currently no unified data platform covering all modalities. Therefore, this study selected the following representative multimodal anger emotion recognition methods for a brief comparison:

CNN+Bi-LSTM+HAM [38]: This method introduces HAM on the basis of the CNN + Bi-LSTM network framework. The mechanism can consider features of different levels and types and adaptively adjust the attention mechanism to selectively focus on key facial features. This multimodal framework combines driver's voice, facial images, and video sequence data for emotion recognition, achieving an anger emotion recognition accuracy of 85%.

CBLNN [40]: This method uses facial geometric features obtained by Convolutional Neural Network (CNN) as intermediate variables for Bidirectional Long Short-Term Memory (Bi-LSTM) heart rate analysis. Subsequently, the output of Bi-LSTM is used as input for the CNN module to extract listening rate features. Finally, Multimodal Factorized Bilinear Pooling (MFB) is used to fuse the extracted information for emotion recognition. Using facial and heart rate data, this dual-modal method achieves an anger emotion recognition accuracy of 90.5%.

AM-LSTM [36]: This study combines LSTM with an attention mechanism for anger emotion recognition through multimodal decision-level fusion of EEG signals, physiological signals, and driving behavior information, achieving an accuracy of 76.3%.

SVM-DS [34]: This study based on ECG signals and driving behavior signals for anger emotion multimodal fusion recognition shows that the SVM-DS model with multimodal decision-level fusion performs best, with an accuracy of 84.8%.

Table 12 lists the results of the proposed method and other advanced models on different datasets in terms of various evaluation metrics. The proposed model (CA-MDER) outperforms all other models in both ACC and F1 metrics. In summary, compared to classic methods, the CA-MDER model proposed in this study demonstrates good classification and generalization capabilities.

TABLE 12. Performance comparison of various multimodal recognition methods.

Model	ACC(%)	F1(%)
CNN+Bi-LSTM+HAM	85.0	83.3
CNLNN	90.5	-
AM-LSTM	76.3	73.2
SVM-DS	84.8	85.0
CA-MDER	91.68	90.37

F. ABLATION EXPERIMENT

To verify the effectiveness of the context-aware module and multimodal fusion, an ablation experiment was designed for the proposed model, including the removal of context awareness and fusion of each single modality data. M represents the multimodal task, F represents the facial single modality, V represents the voice single modality, S represents the driving state single modality, and Awareness represents the context-awareness module. The ablation experiment results for each indicator are shown in Table 13.

TABLE 13. Results of modality ablation experiment: Multimodal dataset.

M	F	V	S	Awareness	ACC(%)	F1(%)
√	×	√	√	√	72.53	72.12
√	√	×	√	√	88.43	86.75
√	√	√	×	√	87.57	87.31
√	√	√	√	×	86.72	84.91
√	√	√	√	√	91.68	90.37

V. RESULTS AND DISCUSSION

To enhance the accuracy of driver anger emotion recognition, this study proposed a context-aware multimodal driver anger emotion recognition method (CA-MDER), integrating facial, voice, and vehicle driving state information. First, anger emotion recognition was conducted for each single modality. To improve recognition accuracy, we initially proposed a facial emotion recognition method based on Attention Mechanism Deep Separable Convolutional Neural Network (AM-DSCNN). This method focuses on key facial features determining emotions using an attention module and then enhances model computational efficiency by introducing deep separable convolution modules. Next, the SVM used for speech emotion recognition was improved by considering the temporal characteristics of voice data and proposing a hybrid

kernel function combining Dynamic Time Warping (DTW) and RBF, effectively handling time elasticity in time series data and improving the optimal correspondence between time series. Finally, for vehicle driving state emotion recognition, features capturing vehicle interactions were introduced, and a driver's driving style recognition module was proposed to mitigate the impact of driving heterogeneity on anger emotion recognition. After completing emotion recognition for each modality, this study used the optimal weight allocation under adaptive scenarios obtained through context awareness for multimodal decision-level fusion, ultimately outputting emotion classification results.

The recognition results show (Table 11) that the CA-MDER model proposed in this paper achieves an accuracy of 91.68% and an F1 score of 90.37%. Compared to other advanced multimodal anger emotion recognition models (Table 12), CA-MDER achieves high levels in both accuracy and F1, surpassing existing multimodal recognition models. Additionally, the ablation experiment revealed some notable points. Although the recognition accuracies of individual modalities vary significantly, the overall recognition rate still increases after multimodal fusion, aligning with current mainstream research and possibly corroborating the view that drivers may hide their emotions in certain modalities under some scenarios. However, in some cases, even using more modalities for fusion may not yield ideal results. For example, in the ablation experiment, directly fusing facial, voice, and driving state modalities without using context-aware adaptive weight allocation only slightly improved recognition rate by about 1% (compared to the highest accuracy of facial single modality, same below). In contrast, using context-aware adaptive weight allocation for fusion of facial and driving state modalities alone can improve recognition accuracy by about 3%. If fusing facial, voice, and driving state modalities, the improvement in recognition accuracy can reach about 5%. This validates the effectiveness of the context-aware adaptive weight allocation method proposed in this study.

Looking forward, under the support of the National Key R&D Program Project ''Major Accident Risk Prevention and Emergency Avoidance Technology for Road Transport Vehicles'' (2023YFC3009600) and the Graduate Innovation Fund of Jilin University, this study will explore applications in accident risk prevention, emergency avoidance, and other aspects related to road transport vehicles. Due to limitations in research duration and measurement methods, this paper has some limitations, and we plan to make improvements in the following areas: a. In this paper, the attention mechanism is only used to focus on facial features. In the future, we will introduce the attention mechanism into other modalities' data to explore possibilities for improving overall recognition accuracy; b. The type and amount of data in real vehicle scenarios is relatively small, and multimodal data can be collected more easily in the future, e.g., through on-board sensors or vehicle networking; c. The scenarios used to train the context-aware adaptive weight allocation method in this paper are somewhat limited. In the future, we will consider enriching the types of scenarios from aspects such as weather and light intensity.

REFERENCES

[1] B. Parkinson, ''Anger on and off the road,'' *Brit. J. Psychol.*, vol. 92, no. 3, pp. 507–526, Aug. 2001, doi: 10.1348/000712601162310.
[2] J. L. Deffenbacher, E. R. Oetting, and R. S. Lynch, ''Development of a driving anger scale,'' *Psychol. Rep.*, vol. 74, no. 1, pp. 83–91, Feb. 1994, doi: 10.2466/pr0.1994.74.1.83.
[3] E. R. Dahlen, R. C. Martin, K. Ragan, and M. M. Kuhlman, ''Driving anger, sensation seeking, impulsiveness, and boredom proneness in the prediction of unsafe driving,'' *Accident Anal. Prevention*, vol. 37, no. 2, pp. 341–348, Mar. 2005, doi: 10.1016/j.aap.2004.10.006.
[4] J. Lu, X. Xie, and R. Zhang, ''Focusing on appraisals: How and why anger and fear influence driving risk perception,'' *J. Saf. Res.*, vol. 45, pp. 65–73, Jun. 2013, doi: 10.1016/j.jsr.2013.01.009.
[5] P. Ekman and W. V. Friesen, ''Facial action coding system,'' *Environ. Psychol. Nonverbal Behav.*, Jan. 1978.
[6] A. Barman and P. Dutta, ''Facial expression recognition using distance and texture signature relevant features,'' *Appl. Soft Comput.*, vol. 77, pp. 88–105, Apr. 2019, doi: 10.1016/j.asoc.2019.01.011.
[7] T. Ojala, M. Pietikäinen, and D. Harwood, ''A comparative study of texture measures with classification based on featured distributions,'' *Pattern Recognit.*, vol. 29, no. 1, pp. 51–59, Jan. 1996.
[8] N. Dalal and B. Triggs, ''Histograms of oriented gradients for human detection,'' in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2005, pp. 886–893, doi: 10.1109/CVPR.2005.177.
[9] H. Ali, M. Hariharan, S. Yaacob, and A. H. Adom, ''Facial emotion recognition using empirical mode decomposition,'' *Expert Syst. Appl.*, vol. 42, no. 3, pp. 1261–1277, Feb. 2015.
[10] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, ''Facial expression recognition from near-infrared videos,'' *Image Vis. Comput.*, vol. 29, no. 9, pp. 607–619, Aug. 2011.
[11] L. Zhang, K. Mistry, M. Jiang, S. C. Neoh, and M. A. Hossain, ''Adaptive facial point detection and emotion recognition for a humanoid robot,'' *Comput. Vis. Image Understand.*, vol. 140, pp. 93–114, Nov. 2015.
[12] Z. Zhang, L. Cui, X. Liu, and T. Zhu, ''Emotion detection using Kinect 3D facial points,'' in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. (WI)*, Oct. 2016, pp. 407–410. Accessed: Nov. 9, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7817080/
[13] R. Jiang, A. T. S. Ho, I. Cheheb, N. Al-Maadeed, S. Al-Maadeed, and A. Bouridan, ''Emotion recognition from scrambled facial images via many graph embedding,'' *Pattern Recognit.*, vol. 67, pp. 245–251, Jul. 2017.
[14] K. Candra Kirana, S. Wibawanto, and H. Wahyu Herwanto, ''Facial emotion recognition based on viola-jones algorithm in the learning environment,'' in *Proc. Int. Seminar Appl. Technol. Inf. Commun.*, Sep. 2018, pp. 406–410. Accessed: Nov. 9, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8549735/
[15] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos, ''Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order,'' *Pattern Recognit.*, vol. 61, pp. 610–628, Jan. 2017.
[16] M. Liu, S. Li, S. Shan, and X. Chen, ''AU-aware deep networks for facial expression recognition,'' in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–6, doi: 10.1109/FG.2013.6553734.
[17] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, ''Deep learning for emotion recognition on small datasets using transfer learning,'' in *Proc. ACM Int. Conf. Multimodal Interact.*, Nov. 2015, pp. 443–449, doi: 10.1145/2818346.2830593.
[18] R. S. John, S. B. Alex, M. S. Sinith, and L. Mary, ''Significance of prosodic features for automatic emotion recognition,'' *AIP Conf. Proc.*, vol. 2222, no. 1, Apr. 2020, Art. no. 030003, doi: 10.1063/5.0004235.
[19] C. Nussbaum, A. Schirmer, and S. R. Schweinberger, ''Contributions of fundamental frequency and timbre to vocal emotion perception and their electrophysiological correlates,'' *Social Cognit. Affect. Neurosci.*, vol. 17, no. 12, pp. 1145–1154, Dec. 2022.
[20] S. Lalitha, D. Geyasruti, R. Narayanan, and S. M, ''Emotion detection using MFCC and cepstrum features,'' *Proc. Comput. Sci.*, vol. 70, pp. 29–35, Jan. 2015.

[21] Y. Zhou, J. Li, Y. Sun, J. Zhang, Y. Yan, and M. Akagi, "A hybrid speech emotion recognition system based on spectral and prosodic features," *IEICE Trans. Inf. Syst.*, vol. E93-D, no. 10, pp. 2813–2821, 2010.

[22] H. Aouani and Y. B. Ayed, "Speech emotion recognition with deep learning," *Proc. Comput. Sci.*, vol. 176, pp. 251–260, Jan. 2020, doi: 10.1016/j.procs.2020.08.027.

[23] T. M. Rajisha, A. P. Sunija, and K. S. Riyas, "Performance analysis of Malayalam language speech emotion recognition system using ANN/SVM," *Proc. Technol.*, vol. 24, pp. 1097–1104, Jan. 2016.

[24] E. M. Albornoz, D. H. Milone, and H. L. Rufiner, "Spoken emotion recognition using hierarchical classifiers," *Comput. Speech Lang.*, vol. 25, no. 3, pp. 556–570, Jul. 2011.

[25] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2203–2213, Dec. 2014.

[26] L. Sun, S. Fu, and F. Wang, "Decision tree SVM model with Fisher feature selection for speech emotion recognition," *EURASIP J. Audio, Speech, Music Process.*, vol. 2019, no. 1, p. 2, Dec. 2019, doi: 10.1186/s13636-018-0145-5.

[27] S. Kwon, "CLSTM: Deep feature-based speech emotion recognition using the hierarchical ConvLSTM network," *Mathematics*, vol. 8, no. 12, p. 2133, Nov. 2020.

[28] H. M. M. Hasan and Md. A. Islam, "Emotion recognition from Bengali speech using RNN modulation-based categorization," in *Proc. 3rd Int. Conf. Smart Syst. Inventive Technol. (ICSSIT)*, Aug. 2020, pp. 1131–1136. Accessed: Dec. 19, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9214196/

[29] H. Lei, "The characteristics of angry driving behaviors and its effects on traffic safety," M.S. thesis, Wuhan Univ. Technol., 2011.

[30] M. Zhong, H. Hong, and Z. Yuan, "Experiment research on influence of angry emotion for driving behaviors," *J. Chongqing Univ. Technol. Natural Sci.*, vol. 25, no. 10, pp. 6–11, 2011.

[31] F. Techer, C. Jallais, Y. Corson, F. Moreau, D. Ndiaye, B. Piechnick, and A. Fort, "Attention and driving performance modulations due to anger state: Contribution of electroencephalographic data," *Neurosci. Lett.*, vol. 636, pp. 134–139, Jan. 2017.

[32] L. Precht, A. Keinath, and J. F. Krems, "Effects of driving anger on driver behavior—Results from naturalistic driving data," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 45, pp. 75–92, Feb. 2017.

[33] S. Shafaei, T. Hacizade, and A. Knoll, "Integration of driver behavior into emotion recognition systems: A preliminary study on steering wheel and vehicle acceleration," in *Proc. Asian Conf. Comput. Vis.*, G. Carneiro and S. You, Eds., Cham, Switzerland: Springer, 2019, pp. 386–401, doi: 10.1007/978-3-030-21074-8_32.

[34] F. Wang, "Research on driver anger emotion recognition method based on multimodal fusion," M.S. thesis, Jilin Univ., 2023.

[35] X. Yu, "Driver's anger recognition method based on human-vehicle environment information fusion," M.S. thesis, Shandong Univ. Technol., 2021.

[36] P. Wang, "Study on multimodal recognition method of Driver's anger emotion and mechanism of driving risk under anger emotion," M.S. thesis, Chongqing Univ., 2020.

[37] J. Tang, Z. Ma, K. Gan, J. Zhang, and Z. Yin, "Hierarchical multimodal-fusion of physiological signals for emotion recognition with scenario adaption and contrastive alignment," *Inf. Fusion*, vol. 103, Mar. 2024, Art. no. 102129, doi: 10.1016/j.inffus.2023.102129.

[38] D. Zhou, Y. Cheng, L. Wen, H. Luo, and Y. Liu, "Drivers' comprehensive emotion recognition based on HAM," *Sensors*, vol. 23, no. 19, p. 8293, Oct. 2023, doi: 10.3390/s23198293.

[39] J. Ni, W. Xie, Y. Liu, J. Zhang, Y. Wan, and H. Ge, "Driver emotion recognition involving multimodal signals: Electrophysiological response, nasal-tip temperature, and vehicle behavior," *J. Transp. Eng., A, Syst.*, vol. 150, no. 1, Jan. 2024, Art. no. 04023125, doi: 10.1061/jtepbs.teeng-7802.

[40] G. Du, Z. Wang, B. Gao, S. Mumtaz, K. M. Abualnaja, and C. Du, "A convolution bidirectional long short-term memory neural network for driver emotion recognition," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 7, pp. 4570–4578, Jul. 2021, doi: 10.1109/TITS.2020.3007357.

[41] W. Li, Y. Cui, Y. Ma, X. Chen, G. Li, G. Zeng, G. Guo, and D. Cao, "A spontaneous driver emotion facial expression (DEFE) dataset for intelligent vehicles: Emotions triggered by video-audio clips in driving scenarios," *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 747–760, Jan. 2023, doi: 10.1109/TAFFC.2021.3063387.

[42] A. J. Kayal and J. Nirmal, "Multilingual vocal emotion recognition and classification using back propagation neural network," in *Proc. Advancement Sci. Technol., 2nd Int. Conf. Commun. Syst. (ICCS)*, Rajasthan, India, 2016, Art. no. 020054, doi: 10.1063/1.4942736.

[43] K. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Commun.*, vol. 40, nos. 1–2, pp. 227–256, Apr. 2003, doi: 10.1016/s0167-6393(02)00084-5.

[44] M. Abdelwahab and C. Busso, "Evaluation of syllable rate estimation in expressive speech and its contribution to emotion recognition," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, South Lake Tahoe, NV, USA, Dec. 2014, pp. 472–477, doi: 10.1109/SLT.2014.7078620.

[45] R. Alcaraz and J. J. Rieta, "A novel application of sample entropy to the electrocardiogram of atrial fibrillation," *Nonlinear Anal., Real World Appl.*, vol. 11, no. 2, pp. 1026–1035, Apr. 2010.

[46] L. Sun, B. Zou, S. Fu, J. Chen, and F. Wang, "Speech emotion recognition based on DNN-decision tree SVM model," *Speech Commun.*, vol. 115, pp. 29–37, Dec. 2019, doi: 10.1016/j.specom.2019.10.004.

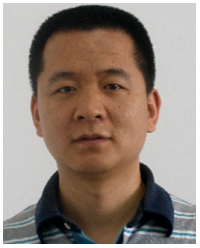
[47] S. Kanwal, S. Asghar, and H. Ali, "Feature selection enhancement and feature space visualization for speech-based emotion recognition," *PeerJ Comput. Sci.*, vol. 8, p. e1091, Nov. 2022, doi: 10.7717/peerj-cs.1091.

[48] T. Zimasa, S. Jamson, and B. Henson, "The influence of driver's mood on car following and glance behaviour: Using cognitive load as an intervention," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 66, pp. 87–100, Oct. 2019, doi: 10.1016/j.trf.2019.08.019.

[49] L. Shamoa-Nir, "Road rage and aggressive driving behaviors: The role of state-trait anxiety and coping strategies," *Transp. Res. Interdiscipl. Perspect.*, vol. 18, Mar. 2023, Art. no. 100780, doi: 10.1016/j.trip.2023.100780.

[50] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Jan. 2019.

[51] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. Schuller, K. Star, E. Hajiyev, and M. Pantic, "SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 1022–1040, Mar. 2021.



TONGQIANG DING received the M.S. and Ph.D. degrees from the School of Transportation, Jilin University, China, in 2001 and 2005, respectively.

He worked at the University of Minnesota, USA, as a Visiting Scholar, in 2014. He is currently an Associate Professor with the School of Transportation, Jilin University, where he is also the Head of the Department of Traffic Engineering, School of Transportation. His research interests include traffic safety of traditional road traffic systems and emerging intelligent transport systems represented by intelligent vehicles and intelligent networks, involving fundamental theories, methods, technologies, and practical applications of traffic safety.

KEXIN ZHANG received the B.S. degree from Shandong University of Science and Technology, in 2022. He is currently pursuing the M.S. degree with the School of Transportation, Jilin University.

His research interests include the analysis and recognition of drivers' psycho-behavioural characteristics and in-vehicle intelligent systems.

SHUAI GAO received the master's degree from the School of Transportation, Jilin University, in 2013.

She became a full-time Faculty Member with the Department of Urban Railway Operation Management, School of Railway and Transportation, Jilin Jiaotong Vocational and Technical College, in 2014, and the Director of the Department of Operation, in 2022. Her research interests include intelligent transport systems, on-board intelligent systems, and traffic microsimulation.

XINNING MIAO received the B.S. degree from Shandong University of Technology, in 2020, and the M.S. degree from the School of Transportation, Jilin University, in 2023.

She is currently working with Beijing Jingwei Hirain Technologies Company Inc. Her research interests include psychological and behavioural characteristics of drivers.

JIANFENG XI received the M.S. and Ph.D. degrees from the School of Transportation, Jilin University, China, in 2003 and 2007, respectively.

He worked at the University of Minnesota, USA, as a Postdoctoral Fellow, in 2010. He became a Professor with the School of Transportation, Jilin University, in 2016. His research interests include driving safety and management and emergency management and simulation systems.