Adv. Manuf.
https://doi.org/10.1007/s40436-024-00519-8

AI-enabled intelligent cockpit proactive affective interaction: middle-level feature fusion dual-branch deep learning network for driver emotion recognition

Ying-Zhang Wu[1] · Wen-Bo Li[1] · Yu-Jing Liu[1] · Guan-Zhong Zeng[2] · Cheng-Mou Li[1] · Hua-Min Jin[3] · Shen Li[4] · Gang Guo[1]

Received: 1 November 2023 / Revised: 23 November 2023 / Accepted: 1 July 2024

Abstract Advances in artificial intelligence (AI) technology are propelling the rapid development of automotive intelligent cockpits. The active perception of driver emotions significantly impacts road traffic safety. Consequently, the development of driver emotion recognition technology is crucial for ensuring driving safety in the advanced driver assistance system (ADAS) of the automotive intelligent cockpit. The ongoing advancements in AI technology offer a compelling avenue for implementing proactive affective interaction technology. This study introduced the multimodal driver emotion recognition network (MDERNet), a dual-branch deep learning network that temporally fused driver facial expression features and driving behavior features for non-contact driver emotion recognition. The proposed model was validated on publicly available datasets such as CK+, RAVDESS, DEAP, and PPB-Emo, recognizing discrete and dimensional emotions. The results indicated that the proposed model demonstrated advanced recognition performance, and ablation experiments confirmed the significance of various model components. The proposed method serves as a fundamental reference for multimodal feature fusion in driver emotion recognition and contributes to the advancement of ADAS within automotive intelligent cockpits.

Keywords Driver emotion · Artificial intelligence (AI) · Facial expression · Driving behavior · Intelligent cockpit

1 Introduction

The automotive intelligent cockpit refers to a mobile space incorporating advanced software and hardware systems, possessing the human-machine-environment fusion capabilities because of human-machine interaction, network-connected services, and scenario expansion, and offering occupants a comprehensive experience of safety, intelligence, efficiency, and pleasure. As a representative cyber-physical-social system (CPSS), it seamlessly integrates diverse technologies, including artificial intelligence (AI), information communication, sensor networks, and augmented reality displays [1, 2]. Intelligent cockpits actively sense the occupants' state, offering an opportunity to address driving safety issues

✉ Wen-Bo Li
wenbo_li@cqu.edu.cn

Ying-Zhang Wu
cquwyz@cqu.edu.cn

Yu-Jing Liu
liuyujing@cqu.edu.cn

Guan-Zhong Zeng
zengguanzhong@hikvision.com

Cheng-Mou Li
cqulcm@cqu.edu.cn

Hua-Min Jin
huaminkim@sina.com

Shen Li
sli299@tsinghua.edu.cn

Gang Guo
guogang@cqu.edu.cn

[1] College of Mechanical and Vehicle Engineering, Chongqing University, Chongqing 400044, People's Republic of China

[2] Hikvision Research Institute, Hangzhou 311599, People's Republic of China

[3] China Society of Automotive Engineers, Beijing 100021, People's Republic of China

[4] Department of Civil Engineering, Tsinghua University, Beijing 100084, People's Republic of China

Published online: 04 September 2024

conventionally attributed to human factors [3]. The advanced driver assistance system (ADAS) within an automotive intelligent cockpit dynamically monitors and intervenes in the driver's state [4]. This system is a prerequisite for ensuring the safe operation of human-machine co-driving vehicles [5]. In addition, it plays a crucial role in mitigating road traffic accidents that occur during the human-machine co-driving stage, thereby enhancing overall road safety [6]. Driver-related factors contributed to over 90% of road traffic accidents [7], with approximately 15% attributed to drivers' emotions and behaviors [8]. The heightened risk associated with driving under the influence of drivers' emotions has emerged as a substantial contributor to road safety hazards [9]. Hence, the advancement of driver emotion recognition technology is paramount for enabling proactive affective interaction of ADAS within automotive intelligent cockpits.

Addressing the impact of driver emotions on road traffic safety necessitates precise, dependable, and efficient recognition of emotional states within automotive intelligent cockpits. Driver emotions exhibit multifaceted variations, encompassing intricate interplays between physiological and behavioral aspects within diverse driving contexts [10–14]. Conventional emotion recognition approaches, relying on facial feature modeling, face inherent challenges in achieving accuracy and practical reliability. The ongoing advancements in AI technology, particularly its robust feature extraction and modeling capabilities [15, 16], offer a compelling avenue for implementing driver emotion recognition technology.

Facial expressions are fundamental manifestations of emotional states and constitute a pivotal route for implementing emotion recognition within AI technology [17]. Nevertheless, within the context of the driving environment, subtle changes in the driver's facial expressions prove more elusive than those encountered in daily life [18, 19]. This inherent subtlety poses significant challenges for vision-based AI emotion recognition techniques. Concurrently, the driver's emotional representation manifests across various dimensions, including driving behavior and physiological signals [20]. Consequently, AI-based multimodal feature fusion techniques hold significant promise in enhancing emotion recognition accuracy [21].

This study introduced the multimodal driver emotion recognition network (MDERNet) for automotive intelligent cockpit to resolve the above mentioned limitations. MDERNet is a dual-branch deep learning architecture illustrated in Fig. 1, which employed facial expression and driving behavior.

The main contributions of this study could be concluded as follows.

(i) A dual-branch driver emotion recognition model named MDERNet was proposed to achieve non-contact dynamic driver emotion recognition with the fusion of facial expression and driving behavior.

(ii) The frame attention and fusion modules in MDERNet facilitated the intermediate fusion process between the facial expression feature extraction branch and the driving behavior feature extraction branch, resulting in enhanced performance for driver emotion recognition.

(iii) The proposed model effectively recognized seven discrete emotions (anger, disgust, sadness, fear, happiness, surprise, and neutral) and three types of dimen-

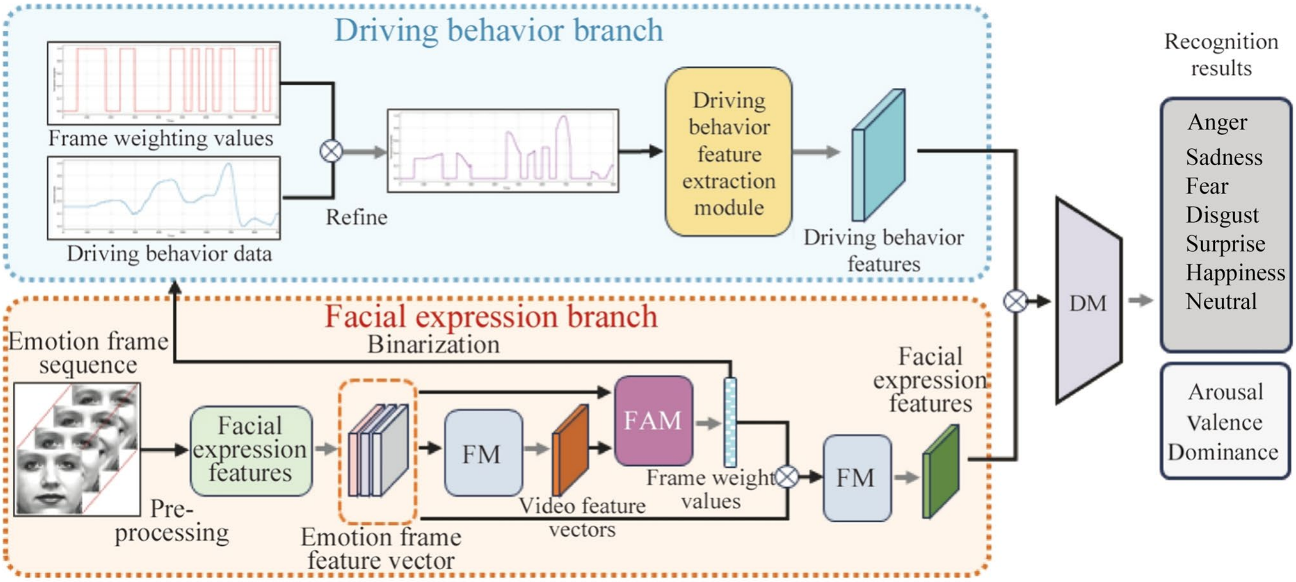


Fig. 1 Architecture of driver emotion recognition model based on facial expression and driving behavior

sional emotions (arousal, valence, and dominance). The results demonstrated MDERNet's strong performance in driver emotion recognition.

The remainder of this study is as follows. Related works about emotion recognition are summarized in Sect. 2. The proposed MDERNet is introduced in detail in Sect. 3. Section 4 introduces the experimental setup and analyzes the results of MDERNet. The conclusion is described in Sect. 5.

2 Related works

2.1 Discrete emotion and dimensional emotion

Emotions can be classified into two main categories: discrete and dimensional, as proposed by various emotion models. Ekman [22] proposed that emotions were discrete, identifying six fundamental emotions: happiness, sadness, anger, fear, surprise, and disgust. In addition, other emotions were regarded as intricate combinations of these foundational states. However, the discrete emotion model relies on verbal descriptions for emotion categorization rather than quantitative analysis. This approach poses challenges in analyzing complex emotions [23].

Lang [24] introduced the "two-dimensional valence-arousal model", which categorized emotions along two dimensions: valence and arousal. Valence represents the degree of pleasure associated with an emotion, while arousal reflects the intensity of the emotional experience. Mehrabian's extension of the emotion model introduces a third dimension—dominance [25], representing the continuum from submissiveness to dominance. This dimension reflects an individual's capacity to regulate emotions. Therefore, the dimensional emotion model quantitatively characterizes emotions across three primary dimensions: valence, arousal, and dominance. Here, we balanced the assessment of model performance between the discrete emotion model and the dimensional emotion model.

2.2 AI-based emotion recognition

During driving, the driver's emotional representation is mainly reflected in the driver's facial expression [26, 27], driving behavior [28], vocal responses [29], and physiological indicators [30]. In AI-based emotion recognition approaches, researchers commonly focus on facial expressions as a primary modality and integrate additional modalities.

Physiological features are suitable for recognizing internal emotions. Wu and Li [18] introduced a multimodal approach for emotion identification that combined facial expression analysis utilizing a multi-level convolutional neural network (CNN) model and electroencephalography (EEG) information with a stacked bidirectional long short term memory (Bi-LSTM) model. The D-S evidence theory was employed at the decision level to fuse the emotion identification results. Ali and Hughes [31] introduced the unified biosensor-vision multimodal transformer-based (UBVMT) method, which classified emotions in an arousal-valence space by integrating a 2D representation of an electrocardiogram/photoplethysmography (ECG/PPG) signal with facial information. However, the method of physiological signal acquisition faces limitations owing to its invasive nature and susceptibility to interference, rendering it less suitable for specific applications.

Extracting speech features for emotion recognition has proven to be an effective method [32]. Liu et al. [33] introduced a multimodal fusion network (M2FNet) that leveraged complementarity and importance for emotion recognition. By accounting for critical differences between multiple modalities (vision and audio), they assign weights through an attention network based on their relative significance. Mocanu et al. [34] proposed an end-to-end multimodal emotion recognition framework incorporating self-attention mechanisms for audio and visual modalities. The fusion strategy involves cross-attention for combining audio and video features. Nevertheless, this method necessitates high-quality speech signals and encounters challenges in noisy driving environments.

The fusion of driver facial expression and driving behavior features exhibited non-invasiveness and stable anti-interference properties, effectively addressing the above-mentioned challenges. In addition, data-level fusion, the middle-level feature fusion and decision-level fusion are widely employed multimodal information fusion ways in AI technologies [18]. However, the middle-level feature fusion is the most effective but challenging. In our proposed method, we accounted for the continuity of emotional expression by fusing facial expression and driving behavior features along the time dimension. This approach maximizes the extraction of the driver's emotional features.

3 Description of MDERNet model

3.1 Overall structure of MDERNet model

The proposed MDERNet model based on facial expression and driving behavior is a two-branch network that combines multimodal data for discrete and dimensional emotion recognition utilizing sparse representations and attention mechanisms, as illustrated in Fig. 1. The two branches of MDERNet handle facial expression modal and driving behavior modal features, respectively. MDERNet utilizes facial expressions to generate temporal attention and driving behavior to refine input features for feature fusion between these two modalities. The two-branch MDERNet model comprises five modules: facial expression feature extraction module (FEFEM), fusion module (FM), frame attention module (FAM), driving behavior feature extraction module (DBFEM), and decision module (DM).

Each MDERNet input sample comprises a sequence of consecutive frame images from a single video and a corresponding numerical sequence of driving behavior data. Firstly, the facial expression branch represents preprocessed video frames as consecutive face images fed into FEFEM, a deep CNN without fully connected or classification layers. Consecutive frames of the same video are sent through FEFEM to obtain primary video-level features in FM. The FAM is the attention module that determines the overall importance of each frame in the video. The FAM module has two inputs: deep features obtained by FEFEM for each frame image, and video-level features obtained by FEFEM and FM for all frame images in that video. FAM's frame attention weights reflect critical deviations in the temporal sequence data. These weights are multiplied with their corresponding deep feature outputs from FEFEM, which are then integrated with other frames' deep features after recalibration. The refined overall features of the video are obtained by inputting these integrated features into FM. Simultaneously, the driving behavior branch receives integrated frame attention weights chronologically. Another branch of MDERNet processes driving behavior data, normalized with a mean of 0 and a variance of 1. Frame attention weights from the facial expression branch are processed, up-sampled to match driving behavior data length, and benchmarked to obtain binarized continuous values. The driving behavior features are obtained by multiplying the binarized continuous values with the driving behavior data individually and inputting them into DBFEM. Finally, the features extracted from the two branches are spliced and inputted into DM, which comprises a conventional multilayer fully connected layer to recognize the driver's emotions based on facial expressions and driving behavior features.

3.2 Facial expression branch

3.2.1 FEFEM

The FEFEM utilizes a convolutional neural network to extract deep features of faces with input facial expressions. Because ResNet is currently the most widely utilized CNN feature extraction network, this paper utilizes ResNet18 to extract expression features. The input to the FEFEM is a 112×112 resolution grayscale image I_{frame} of a face, and the output is a 512-dimensional 1×1 feature map $\boldsymbol{M}_{\text{fefem}}$, which proceeds as

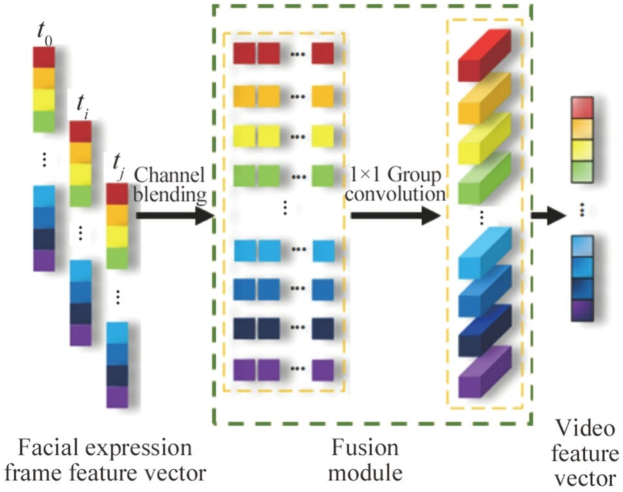


Fig. 2 Fusion module

$$\boldsymbol{M}_{\text{fefem}} = f_{\text{fefem}}(I_{\text{fefem}}), \tag{1}$$

where f_{fefem} is a function of the FEFEM, and $\boldsymbol{M}_{\text{fefem}}$ represents the feature vector for each frame of the face image output. All frame feature vectors belonging to the same video are input into FM and utilized as part of FAM input.

3.2.2 FM

The FM module is illustrated in Fig. 2. The input to FM is a set of feature vectors obtained from all video frame images through FEFEM. FM comprises two steps: channel blending and 1×1 convolution. Inspired by ShuffleNet [35], the Channel blending operation arranges face feature vectors channel-by-channel in a time sequence such that all frame feature maps of each channel can be conveniently convolved in a 1×1 group to form a video feature map by channel. All video feature maps of all channels form the overall features of the video. Feature vectors extracted by FM as overall video feature vectors are input into FAM along with frame feature vectors output from FEFEM. The process is expressed as

$$M_{\text{fm}} = f_{\text{fm}}(M_{\text{fefem}_0}, \cdots, M_{\text{fefem}_{k-1}}), \tag{2}$$

where f_{fm} is a function of FM, including channel mixing and group convolution operations, and k is the number of sampled frames of the video. M_{fefem_i} is the feature mapping obtained from the i-th image frame after FEFEM, and M_{fm} is the overall feature vector of the video after FM processing.

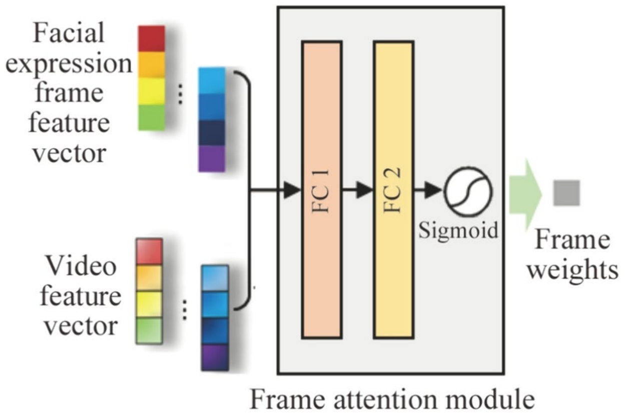


Fig. 3 Frame attention module

3.2.3 FAM

As illustrated in Fig. 3, the FAM is similar to the SE module [36], including two fully connected layers and the Sigmoid function, and the FAM is implemented as

$$\begin{cases} W_{\text{fam}_i} = f_{\text{fam}}\left(\text{Concat}\left(M_{\text{fefem}_i}, M_{\text{fm}}\right)\right), \\ M_{\text{fam}_i} = W_{\text{fam}_i} \times M_{\text{fefem}_i}, i \in [0, k-1], \end{cases} \tag{3}$$

where f_{fam} is the module of FAM, M_{fefem_i} the feature mapping of the i-th frame image obtained by FEFEM, M_{fm} the preliminary video feature mapping of the video to which the i-th frame image belongs, the generated W_{fam_i} denoted as the weight value of the i-th frame image, and M_{fam_i} the feature mapping of the i-th frame image after weighting.

The weight sequence generated by FAM serves two purposes: one is to calibrate deep features M_{fefem_i} across all image frames with $W_{\text{fam}_i} \times M_{\text{fefem}_i}$, and the other is to refine driving behavior data through another branch of the multi-modal model.

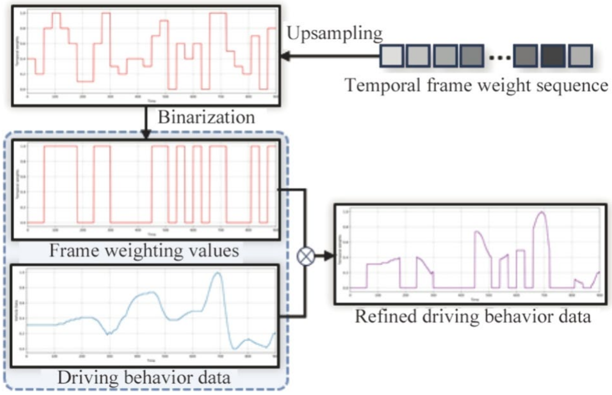


Fig. 4 Process of driving behavior data refine

3.3 Driving behavior branch

3.3.1 Driving behavior data refinement module

Driving behavior data refinement involves filtering and highlighting data from the driving behavior modality with temporal attention obtained from the facial expression modality to achieve information fusion between multiple modalities at the input information level.

The filtering and highlighting of driving behavior data with W_{fam_i} involves two main steps, as illustrated in Fig. 4. Firstly, W_{fam_i} is up-sampled to match the numerical length of driving behavior data I_{db}. Next, W_{fam_i} is binarized by applying a threshold, and the resulting binary values are multiplied with the driving behavior data individually. The implementation process is shown as

$$\begin{cases} W_{\text{fam}}^{\text{binary}} = \text{Binary}\left(\text{Upsample}\left(W_{\text{fam}}\right)\right), \\ I_{\text{db}}^{\text{refined}} = W_{\text{fam}}^{\text{binary}} \times I_{\text{db}}, \end{cases} \tag{4}$$

where Binary and Upsample denote a binarization operation and an upsampling operation, $W_{\text{fam}}^{\text{binary}}$ a sequence of temporal frame attentional weights after upsampling and binarization, and $I_{\text{db}}^{\text{refined}}$ the refined driving behavior data.

3.3.2 DBFEM

DBFEM is a multilayer perceptual machine. The input comprises refined driving behavior data that include steering wheel rotational speed, accelerator pedal angle, brake pedal force, longitudinal velocity, longitudinal acceleration, lateral velocity, and lateral acceleration. Firstly, normalized preprocessing is performed on the input data to ensure consistency in scale. Next, the temporal frame attention weight sequence filters and highlights the sampled driving behavior data. Finally, the selected driving behavior data are spliced together and fed into DBFEM for processing as

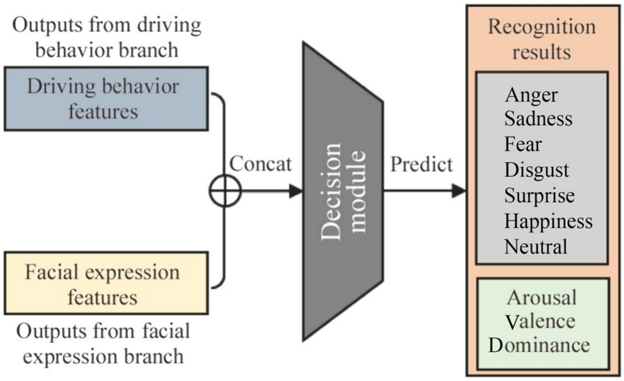

Outputs from driving behavior branch
Driving behavior features
Concat
Decision module
Predict
Facial expression features
Outputs from facial expression branch
Recognition results
Anger
Sadness
Fear
Disgust
Surprise
Happiness
Neutral
Arousal
Valence
Dominance

Fig. 5 Decision module

$$M_{\text{dbfem}} = f_{\text{dbfem}}\left(I_{\text{db}}^{\text{refined}}\right), \tag{5}$$

where f_{dbfem} is a function of DBFEM, and M_{dbfem} is the extracted driving behavioral features.

3.4 DM

The DM is a one-layer, fully connected layer with an attached Softmax function for discrete emotion classification. The driving behavior features and facial expression features extracted from the two branches of MDERNet serve as inputs, as illustrated in Fig. 5. The facial expression feature is obtained by passing the weighted frame image feature M_{fam_i} through another FM with the following process

$$\hat{y} = f_{\text{dm}}\left(\text{Concat}\left(M_{\text{dbfem}}, f_{\text{fm}}\left(M_{\text{fam}_0}, \cdots, M_{\text{fam}_{k-1}}\right)\right)\right), \tag{6}$$

where f_{dm} is a function of DM, and $\hat{y}$ is the model prediction result.

3.5 Loss function

This paper employed different loss functions to evaluate various emotion recognition metrics. Specifically, the cross-entropy [37] loss function was utilized to measure. Accuracy, the F1 loss function [38] was utilized to calculate F1-score, and the mean square error (MSE) loss function [39] was utilized to quantify MSE. The consistency correlation coefficient (CCC) loss function [40] assessed the CCC. The corresponding formulas are expressed as

$$L_{\text{cross entrophy loss}} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{e^{h_{yi}}}{\sum_{j=1}^{C} e^{h_j}}\right), \tag{7}$$

$$L_{F_1 \text{ loss}} = 1 - 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}, \tag{8}$$

$$L_{\text{MSE loss}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{M}\left(I_t^i\right)^2, \tag{9}$$

$$L_{\text{CCC loss}} = 1 - \frac{2S_{\text{C}}}{S^2 + \hat{S}^2 + (\bar{y} - \bar{\hat{y}})^2}, \tag{10}$$

where x_i represents the i-th sample's input feature in the final classification layer. The true label and predicted label of the i-th sample are represented by $y_i \in \{1, 2, \cdots, C\}$ and $\hat{y}_i \in \{1, 2, \cdots, C\}$, respectively. The corresponding averages for y_i and $\hat{y}_i$ are represented by $\bar{y}$ and $\bar{\hat{y}}$. The variances for y_i and $\hat{y}_i$ are represented by S and $\hat{S}$, respectively. S_{C} represents the covariance between y_i and $\hat{y}_i$. The network's output for recognition of the i-th sample is represented by $\boldsymbol{h} = \left(h_1, h_2, \cdots, h_C\right)^{\text{T}}$, where C represents the number of classes.

4 Experimental setup and results

4.1 Data utilized

Because the sample sizes of existing facial expression datasets are generally small, the MS-Celeb-1M face dataset [41] is utilized to pre-train the FEFEM. This pre-training endows the FEFEM with more robust feature extraction and expression recognition capabilities.

This paper validates the effectiveness of the MDERNet model from two perspectives: the discrete sentiment model and the dimensional sentiment model. CK+ and RAVDESS datasets with discrete sentiment labels were selected to validate their performance on discrete sentiment classification tasks. The DEAP dataset with dimensional sentiment labels was selected to validate its performance on the dimensional sentiment regression task. The PPB-Emo dataset was utilized for both classification and regression tasks to validate the effectiveness of the MDERNet model, as it included data on driving behavior in addition to discrete and dimensional sentiment labels.

CK+ [42]. CK+ is a dataset that contains 593 video sequences of spontaneous and performs facial expressions of emotions from 123 participants, along with other metadata. The participants were predominantly female and aged between 18 and 30 years old. Out of the 327 video sequences from 118 participants, seven discrete emotions were labeled: anger, disgust, sadness, fear, happiness, surprise, and contempt. The image sequences had 640 × 480 and 640 × 490 pixel resolutions.

RAVDESS [43]. The RAVDESS dataset comprises 7 356 audiovisual files of emotional speech and singing performances captured by 24 professional actors (12 female, 12 male). The video-voice files are labeled with eight categories of emotions: calm, happy, sad, angry, fearful, surprised, disgusted, and neutral. Each category comprises two emotional intensities: normal and strong. The videos have a pixel resolution of 1 280 × 720.

DEAP [44]. The DEAP dataset contains physiological signals (peripheral physiological data, EEG data, and frontal face data) from 32 participants. Each participant rated their arousal, dominance, and valence on a 9-point scale based on the dimensional emotion model. Each participant watched 40 1-min-long music video elicitation materials, resulting in 880 facial video sequences with a pixel resolution of 720 × 576.

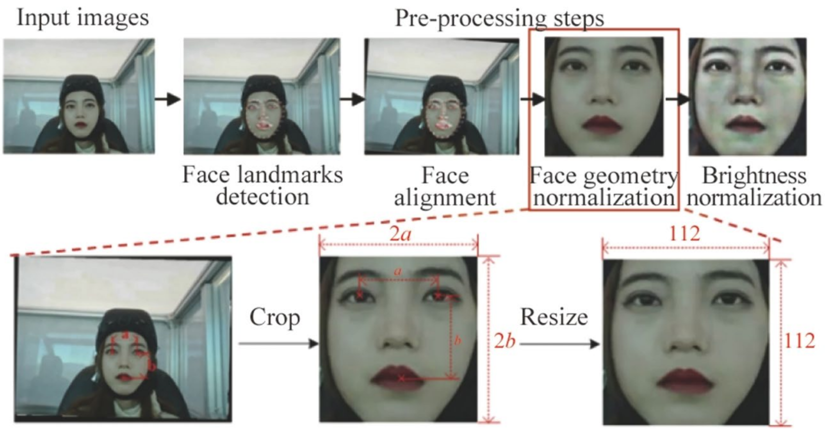


Fig. 6 Process of facial expression preprocessing

PPB-Emo [45]. The PPB-Emo dataset comprises 280 facial video sequences and driving behavior sequences from 40 participants who were successfully emotionally induced. The emotions were labeled according to seven categories of discrete emotions (anger, disgust, sadness, fear, happiness, surprise, and neutral) and the dimensional emotions of arousal-valence-dominance. The face expression video sequences and driving behavior data from 15 s following the start of driving were processed as valid data. The videos have a pixel resolution of 640 × 480.

4.2 Data preprocessing

4.2.1 Facial expression sequence data

The preprocessing of face images mainly involves detecting key points, aligning faces, geometrically normalizing face images, and normalizing brightness. The preprocessing process is presented in Fig. 6.

Firstly, the face image was recognized with the multitask cascade convolutional network (MTCNN) [46] to obtain 68 face key points. Secondly, the angle between the line connecting the left eye center and the right eye center and the horizontal line was then calculated, and the image was rotated such that the line connecting the left eye center and the right eye center was horizontal to achieve face alignment. The distance between the center of the left eye and the center of the right eye is represented by a, and the vertical difference between the center of the left eye (or right eye) and the center of the mouth is represented by b. Finally, the face image was cropped to $2a$ width and $2b$ height before being scaled to an image of 112 × 112 pixels, as illustrated in Fig. 6.

Geometric normalization enables the same facial feature points to coexist roughly in the exact same location in different video frames. This process also discards background details and facial regions, such as ears and forehead, which are unrelated to facial expressions, as they do not represent expression-specific information. The luminance of cropped facial images was normalized to reduce the impact of illumination changes on image signals. We sampled 30 frames from each video for each dataset as training samples for the model.

4.2.2 Driving behavior data

This study selected seven types of driving behavior data: steering wheel rotational speed, accelerator pedal angle, brake pedal force, longitudinal velocity, longitudinal acceleration, lateral velocity, and lateral acceleration. The selected driving behavior data were normalized as

$$x' = \frac{x - \mu}{\sigma}, \tag{11}$$

where x is the raw driving behavior data, u the mean of all data of a particular type, σ the variance of all data of that type, and x' the normalized data.

4.3 Experiment details and evaluation metrics

4.3.1 Experiment details

This study's hardware configuration utilized for model training and testing comprised an NVIDIA Tesla V100 GPU with 32 GB of video memory, running on the Ubuntu 18.04 operating system. The deep learning framework utilized was PyTorch 1.5. During training, we utilized the stochastic gradient descent (SGD) optimizer with a Nesterov momentum of 0.9 and a batch size of 64. The learning rate decay strategy was cosine decay, and the learning rate hot restart epoch was set to 5.

To ensure that the tasks were independent, i.e., no one's video sequences would be present in two-fold and above data segmentation, and all datasets utilized for recognition training here were constructed as training/testing sets by 10-fold cross-validation. The datasets were ordered by participant number and sampled at intervals of 10 to create a ten-fold data subset. The final experimental results in this paper were the average results obtained from ten-fold cross-validation.

4.3.2 Evaluation metrics

This study employed both discrete and dimensional emotion models to quantify emotions. The discrete emotion model categories included anger, disgust, fear, happiness, neutral, sadness, and surprise. Positive real values for the arousal, valence, and dominance dimensions characterized the dimensional emotion model. The discrete emotion and dimensional emotion recognition tasks corresponded to the classification and regression tasks in deep learning, respectively. Accuracy and F1 score were utilized in the classification task as evaluation metrics of model performance, while the regression task utilized mean square error (MSE) and concordance correlation coefficient (CCC) [47]. MSE was utilized to measure the overall mean deviation between the true value θ and its estimate $\hat{\theta}$. The smaller the MSE, the better the model performance. In addition, CCC, whose value ranges from −1 (completely inconsistent) to 1 (completely consistent), was utilized to measure the consistency between real and predicted emotions.

4.4 Experimental results and discussion

4.4.1 Facial expression branch ablation experiment

As previously mentioned, the facial expression branch of the MDERNet model includes FEFEM, FM, and FAM. FEFEM is a ResNet18 model without the classification layer and is widely utilized in various feature extraction tasks. To verify the effectiveness of the proposed FM and FAM for facial expression-based emotion recognition, this study designed the facial expression branch ablation experiment for discrete and dimensional emotion recognition. The model was evaluated by including or removing FAM and FM in the facial expression branch. The prediction results of all frame images were averaged

Table 1 Discrete emotion recognition experimental results of facial expression branch

Models	CK+		RAVDESS		PPB-Emo	
	Accuracy	F1 score	Accuracy	F1 score	Accuracy	F1 score
FEB	0.895 1	0.867 3	0.628 1	0.611 8	0.343 2	0.319 2
FEB (w/o FAM)	0.875 0	0.845 2	0.580 2	0.572 5	0.316 1	0.278 1
FEB (w/o FAM/FM)	0.835 2	0.808 3	0.540 6	0.544 8	0.278 6	0.245 9

Note: FEB represents the facial expression feature extraction branch with FAM and FM. FEB (w/o FAM) represents only the facial expression feature extraction branch without FAM. FEB (w/o FAM/FM) represents the facial expression feature extraction branch without FAM and FM.

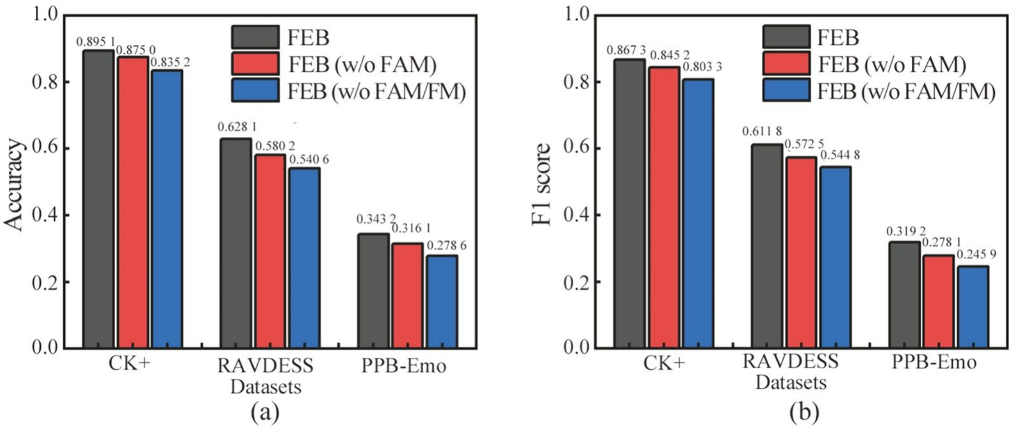


Fig. 7 Discrete emotion recognition experimental results of facial expression model (FEB represents the facial expression feature extraction branch with FAM and FM. FEB (w/o FAM) represents the facial expression feature extraction branch without FAM only. FEB (w/o FAM/FM) represents the facial expression feature extraction branch without FAM and FM.)

as the final result for the facial expression branch model without FAM and FM.

4.4.1.1 Discrete emotion Table 1 presents the results of discrete emotion recognition ablation experiments conducted on the CK+, RAVDESS, and PPB-Emo datasets. The accuracy and F1 score outcomes are presented in Figs. 7a and b, respectively. Higher accuracy and F1 scores correspond to superior performance. Table 1 presents the facial expression feature extraction branch with FAM and FM (FEB) performed best in CK+, RAVDESS, and PPB-Emo, and the facial expression feature extraction branch without FAM only (FEB (w/o FAM)) performed second best. Moreover, the facial expression feature extraction branch without FAM and FM (FEB (w/o FAM/FM)) performed the worst. On CK+, the FEB accuracy (89.51%) is 2.01% and 5.99% higher than the FEB (w/o FAM) (87.50%) and FEB (w/o FAM/FM) (83.52%), respectively. The FEB F1 score (86.73%) is 2.11% and 5.90% higher than the FEB (w/o FAM) (84.52%) and FEB (w/o FAM/FM) (80.83%), respectively. On RAVDESS, the FEB accuracy (62.81%) is 4.79% and 8.75% higher than the FEB (w/o FAM) (58.02%) and FEB (w/o FAM/FM) (54.06%), respectively. The FEB F1 score (61.18%) is 3.93% and 6.70% higher than the FEB (w/o FAM) (57.25%) and FEB (w/o FAM/FM) (54.48%), respectively. On PPB-Emo, the FEB accuracy (34.32%) is 2.71% and 6.46% higher than the FEB (w/o FAM) (31.61%) and FEB (w/o FAM/FM) (27.86%), respectively. The FEB F1 score (31.92%) is 4.11% and 7.33% higher than the FEB (w/o FAM) (27.81%) and FEB (w/o FAM/FM) (24.59%), respectively. The effectiveness of our proposed FEB in discrete emotion recognition was proved.

Table 2 Dimensional emotion recognition experimental results of facial expression branch

Models	DEAP		PPB-Emo	
	MSE	CCC	MSE	CCC
FEB	3.708 3	0.180 0	4.847 8	0.219 0
FEB (w/o FAM)	4.413 5	0.137 8	5.503 7	0.197 0
FEB (w/o FAM/FM)	7.956 5	0.108 5	8.199 1	0.169 8

Note: FEB represents the facial expression feature extraction branch with FAM and FM (FEB (w/o FAM) represents only facial expression feature extraction branch without FAM. FEB (w/o FAM/FM) represents the facial expression feature extraction branch without FAM and FM.)

4.4.1.2 Dimensional emotion Table 2 presents the results of dimensional emotion recognition ablation experiments conducted on the DEAP and PPB-Emo datasets. The MSE and CCC outcomes are illustrated in Figs. 8a and b, respectively. Lower MSE and higher CCC correspond to superior performance. Table 2 presents the FEB that performed best in DEAP and PPB-Emo and the FEB (w/o FAM) that performed second best. Moreover, the FEB (w/o FAM/FM) performed the worst. On DEAP, the FEB MSE (3.708 3) is 0.705 2 and 4.248 2 less than the FEB (w/o FAM) (4.413 5) and FEB (w/o FAM/FM) (7.956 5) respectively. The FEB CCC (0.180 0) is 0.042 2 and 0.071 5 higher than the FEB (w/o FAM) (0.137 8) and FEB (w/o FAM/FM) (0.108 5) respectively. On PPB-Emo, the FEB MSE (4.8478) is 0.6559 and 3.351 3 less than the FEB (w/o FAM) (5.503 7) and FEB (w/o FAM/FM) (8.199 1) respectively. The FEB CCC score (0.219 0) is 0.022 0 and 0.049 2 higher than the FEB (w/o FAM) (0.197 0) and FEB (w/o FAM/FM) (0.169 8), respectively. The effectiveness of our proposed FEB in dimensional emotion recognition was proved.

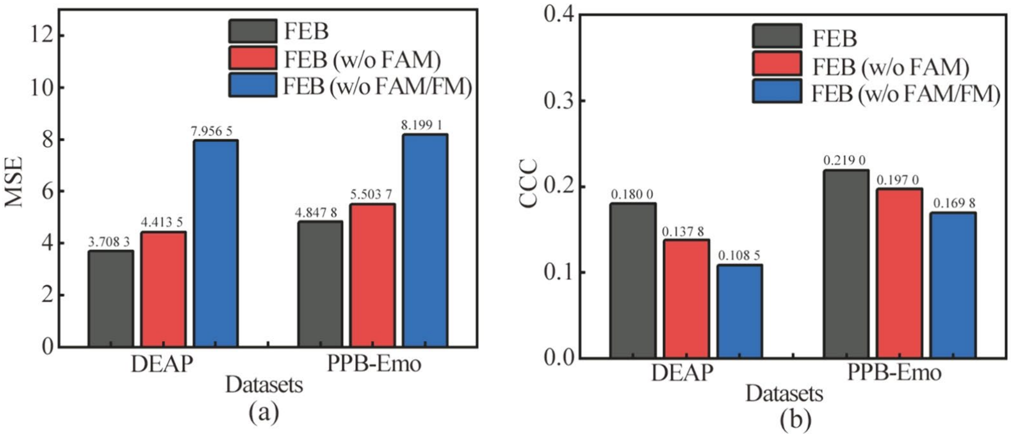


Fig. 8 Dimensional emotion recognition experimental results of facial expression model (FEB represents the facial expression feature extraction branch with FAM and FM. FEB (w/o FAM) represents the facial expression feature extraction branch without FAM only. FEB (w/o FAM/FM) represents the facial expression feature extraction branch without FAM and FM.)

The facial expression branch exhibited optimal performance when incorporating FAM and FM across all datasets (see Tables 1 and 2). Subsequent removal of the FAM module resulted in a consistent decline across all evaluated metrics. Thus, the absence of FAM in the current network architecture prevented direct emphasis on the features of the peak expression frame image, resulting in performance degradation. The model's performance degrades significantly when the facial expression branch omits FAM and FM. This observation underscores the importance of fusing deeper features extracted from all frames of facial expressions for accurate video emotion recognition. The results demonstrate the efficacy of the proposed FAM and FM in emotion recognition from facial expression video data.

4.4.2 *Multimodal feature fusion recognition results of MDERNet model*

The multimodal fusion in the MDERNet model comprised a DBFEM module and a refinement processing module for driving behavior data. The DBFEM was a multilayer perceptual machine, while the refinement processing module comprised a sequence of temporal frame attention weights derived from branching facial expressions. These attention weights served to filter and emphasize relevant driving behavior data. Therefore, this study employed the model without the refinement processing module as the multimodal baseline. This approach allowed us to assess the efficacy of the refinement module in emotion recognition. Notably, for a comprehensive assessment of MDERNet's performance, the complete facial expression branch containing FAM and FM modules was utilized in this experiment.

Table 3 Discrete and dimensional emotion recognition experimental results of MDERNet

Models	PPB-Emo			
	Accuracy	F1 score	MSE	CCC
MDERNet	0.416 7	0.353 1	4.647 5	0.266 1
MDERNet (w/o refine)	0.380 7	0.346 1	4.791 1	0.245 5

Note: MDERNet (w/o refine) represents the MDERNet model without the refinement module.

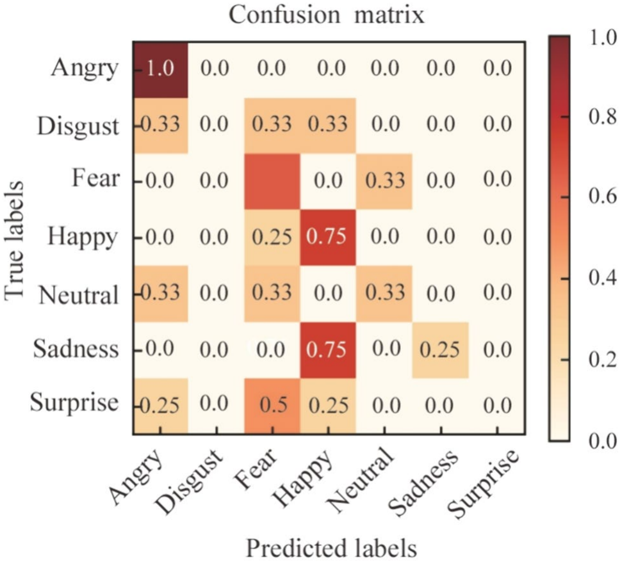


Fig. 10 Confusion matrix of discrete emotion recognition results (MDERNet)

Table 3 presents the experimental results for discrete and dimensional emotion recognition utilizing the MDERNet model in the context of multimodal recognition on the PPB-Emo dataset. Figure 9 presents the performance

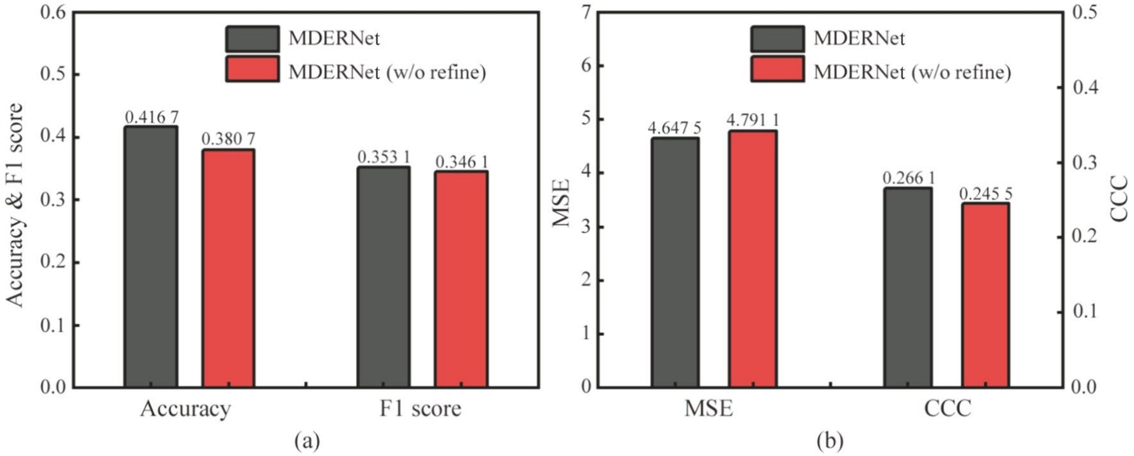


Fig. 9 Discrete and dimensional emotion recognition experimental results of MDERNet (MDERNet (w/o refine) represents the MDERNet model without the refinement module.)

comparison between the complete MDERNet model (indicated by yellow bars) and the MDERNet model without the refinement module (represented by blue bars as MDERNet (w/o refine)). The accuracy and F1 score metrics for discrete emotion recognition are presented in Fig. 9a. In addition, the MSE and CCC metrics for dimensional emotion recognition results are presented in Fig. 9b. The confusion matrix for recognizing discrete emotions utilizing the MDERNet model is demonstrated in Fig. 10. Table 3 presents that the MDERNet performed better than MDERNet (w/o refine) in discrete and dimensional emotion recognition. As for discrete emotion recognition, the MDERNet accuracy (41.67%) is 3.60% higher than the MDERNet (w/o refine) (38.07%). The MDERNet F1 score (35.31%) is 0.70% higher than the MDERNet (w/o refine) (34.61%). As for dimensional emotion recognition, the MDERNet MSE (4.647 5) is 0.143 6 less than the MDERNet (w/o refine) (4.791 1). The MDERNet CCC (0.266 1) is 0.020 6 higher than the MDERNet (w/o refine) (0.245 5). The effectiveness of our proposed MDERNet in discrete and dimensional emotion recognition was proved.

4.4.3 Discussion

Comparing Tables 1, 2, and 3, we observed that the MDERNet model, which incorporated both facial expression and driving behavior modal data, outperformed the facial expression branch containing only facial expression data. This finding underscored the effectiveness of adding the driving behavior modality for accurate emotion recognition. Table 3 revealed that including the refinement module in the driving behavior branch led to significant improvement across all performance indicators for the MDERNet model. The results indicated that enhancing feature fusion between two modalities was more effective when temporal attention was generated through the facial expression modality. In addition, filtering and highlighting input data from the driving behavior modality contributed to refining input features. The efficacy of the proposed modules in the MDERNet model was empirically verified.

Regarding the CK+, RAVDESS, and DEAP datasets, the performance of the facial expression branch within the proposed MDERNet model fell short of achieving the optimal results observed in prior studies [48–50]. This discrepancy primarily arises from variations in data processing techniques and sampling methodologies. As illustrated in Fig. 10, the MDERNet model performed well in identifying drivers' angry, fear, and happy emotions in the PPB-Emo dataset, effectively distinguishing between positive and negative emotions. However, the recognition results for disgust, neutrality, sadness, and surprise emotions were poor. The impact of driving tasks on drivers' facial expressions and driving behavior might explain this phenomenon [51]. Overall, the multimodal driver emotion accuracy based on facial expression and driving behavior (PPB-Emo dataset) remained suboptimal compared to the emotion recognition results obtained from facial expression data in real-life scenarios (CK+, RAVDESS, and DEAP datasets). Consequently, future research should incorporate additional multimodal data to advance our understanding of driver emotion recognition.

5 Conclusions

This study established a multimodal driver emotion recognition model based on driver facial expressions and driving behavior. MDERNet was a deep learning network that achieved the fusion of driver facial expressions and driving behavior for emotion recognition. The model's performance was validated utilizing both the discrete and dimensional emotion models, and the generalizability and sophistication of the model were also validated on other publicly available datasets. The validity of the proposed model architecture and the importance of multimodal fusion in driver emotion recognition methods were verified through branch ablation experiments. The results indicate that MDERNet detection architecture can achieve good detection results for different databases on discrete and dimensional emotion models, respectively. Our proposed method achieves non-contact dynamic driver multiple emotion recognition. The results demonstrated that MDERNet effectively detected emotions across different databases on discrete and dimensional emotion models. In addition, our proposed method achieved non-contact dynamic driver emotion recognition. It serves as a fundamental reference for multimodal feature fusion in driver emotion recognition, contributing to ADAS's advancement within automotive intelligent cockpits.

While this article successfully validates the efficacy of the proposed MDERNet model in identifying driver emotions, it is essential to acknowledge certain limitations. Specifically, the accuracy of driver emotion recognition remains suboptimal owing to the intricate nature of the driving environment. Subsequent research endeavors should consider incorporating additional features influencing driver emotions, including driver attributes (such as driving experience, gender, and age) and physiological signals captured by wearable devices (such as photoplethysmography). These augmentations are expected to enhance the accuracy of driver emotion recognition.

Acknowledgments This work is supported by the National Natural Science Foundation of China (Grant No. 52302497).

References

1. Li W, Wu L, Wang C et al (2023) Intelligent cockpit for intelligent vehicle in metaverse: a case study of empathetic auditory regulation of human emotion. IEEE Trans Syst Man Cybern Syst 53(4):2173–2187
2. Zhao Y, Tian W, Cheng H (2022) Pyramid Bayesian method for model uncertainty evaluation of semantic segmentation in autonomous driving. Automot Innov 5:70–78
3. Zeng X, Wang F, Wang B et al (2022) In-vehicle sensing for smart cars. IEEE Open J Veh Technol 3:221–242
4. Greenwood PM, Lenneman JK, Baldwin CL (2022) Advanced driver assistance systems (ADAS): demographics, preferred sources of information, and accuracy of ADAS knowledge. Transp Res Pt F Traffic Psychol Behav 86:131–150
5. Zhang W, Tang J (2022) Technology developing state and trend about advanced driving assistance system and calculating chip. In: The 4th international academic exchange conference on science and technology innovation (IAECST), Guangzhou, 9–11 Dec, pp 938–943. https://doi.org/10.1109/IAECST57965.2022.10061965
6. Tan Z, Dai N, Su Y et al (2021) Human-machine interaction in intelligent and connected vehicles: a review of status quo, issues, and opportunities. IEEE Trans Intell Transp Syst 23:13954–13975
7. World Health Organization (2018) Global status report on road safety 2018: summary. World Health Organization
8. Ministry of Public Security of the People's Republic of China (2020) One person dies in a car accident every 8 minutes! The highest rate of traffic accidents are these behaviors. http://www.xinhuanet.com/politics/2020-12/02/c_1126809938.htm
9. Quante L, Zhang M, Preuk K et al (2021) Human performance in critical scenarios as a benchmark for highly automated vehicles. Automot Innov 4:274–283
10. Pace-Schott EF, Amole MC, Aue T et al (2019) Physiological feelings. Neurosci Biobehav Rev 103:267–304
11. Adolphs R, Anderson D (2018) The neuroscience of emotion: a new synthesis. Princeton University Press, Princeton
12. Hu H, Zhu Z, Gao Z et al (2018) Analysis on biosignal characteristics to evaluate road rage of younger drivers: a driving simulator study. In: 2018 IEEE intelligent vehicles symposium (IV), 26–30 June, Changshu, pp 156–161
13. Bethge D, Kosch T, Grosse-Puppendahl T et al (2021) Vemotion: using driving context for indirect emotion prediction in real-time. In: The 34th annual ACM symposium on user interface software and technology, 10–13 Oct, pp 638–651
14. Wu X, Wang Y, Peng Z et al (2018) A questionnaire survey on road rage and anger-provoking situations in China. Accid Anal Prev 111:210–221
15. Chen G, Chen K, Zhang L et al (2021) VCANet: vanishing-point-guided context-aware network for small road object detection. Automot Innov 4:400–412
16. Tian C, Leng B, Hou X et al (2022) Robust identification of road surface condition based on ego-vehicle trajectory reckoning. Automot Innov 5:376–387
17. Huang TR, Hsu SM, Fu LC (2021) Data augmentation via face morphing for recognizing intensities of facial emotions. IEEE Trans Affect Comput 14:1228–1235
18. Wu Y, Li J (2023) Multimodal emotion identification fusing facial expression and EEG. Multimed Tools Appl 82:10901–10919
19. Barrett LF, Adolphs R, Marsella S et al (2019) Emotional expressions reconsidered: challenges to inferring emotion from human facial movements. Psychol Sci Public Interest 20:1–68
20. Wang X, Liu Y, Wang F et al (2019) Feature extraction and dynamic identification of drivers' emotions. Transp Res Pt F Traffic Psychol Behav 62:175–191
21. Zhang X, Liu J, Shen J et al (2020) Emotion recognition from multimodal physiological signals using a regularized deep fusion of kernel machine. IEEE T Cybern 51:4386–4399
22. Ekman P (1992) An argument for basic emotions. Cognit Emot 6:169–200
23. Shu L, Xie J, Yang M et al (2018) A review of emotion recognition using physiological signals. Sensors 18:2074. https://doi.org/10.3390/s18072074
24. Lang PJ (1995) The emotion probe: studies of motivation and attention. Am Psychol 50:372. https://doi.org/10.1037/0003-066X.50.5.372
25. Mehrabian A (1996) Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in temperament. Curr Psychol 14:261–292
26. Ekman P, Oster H (1979) Facial expressions of emotion. Annu Rev Psychol 30:527–554
27. Russell JA, Bachorowski JA, Fernández-Dols JM (2003) Facial and vocal expressions of emotion. Annu Rev Psychol 54:329–349
28. Shiota M, Kalat J (2011) Emotion (2nd eds). Wadsworth Cengage Learning Belmont, Australia
29. Bachorowski JA, Owren MJ (2008) Vocal expressions of emotion. Handb Emot 3:196–210
30. Rani P, Liu C, Sarkar N et al (2006) An empirical study of machine learning techniques for affect recognition in human-robot interaction. Pattern Anal Appl 9:58–69
31. Ali K, Hughes CE (2023) A unified transformer-based network for multimodal emotion recognition. arXiv preprint arXiv:230814160. https://doi.org/10.48550/arXiv.2308.14160
32. Li W, Xue J, Tan R et al (2023) Global-local-feature-fused driver speech emotion detection for intelligent cockpit in automated driving. IEEE Trans Intell Veh 8:2684–2697
33. Liu S, Gao P, Li Y et al (2023) Multimodal fusion network with complementarity and importance for emotion recognition. Inf Sci 619:679–694
34. Mocanu B, Tapu R, Zaharia T (2023) Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning. Image Vis Comput 133:104676. https://doi.org/10.1016/j.imavis.2023.104676
35. Zhang X, Zhou X, Lin M et al (2018) Shufflenet: an extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 18–23 June, Salt Lake City, pp 6848–6856
36. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 18–23 June, Salt Lake City, pp 7132–7141
37. Zhang Z, Sabuncu M (2018) Generalized cross entropy loss for training deep neural networks with noisy labels. In: The 32nd conference on neural information processing systems. https://doi.org/10.48550/arXiv.1805.07836
38. Chicco D, Jurman G (2020) The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 21:1–13
39. Rao CR (1980) Some comments on the minimum mean square error as a criterion of estimation. Statistics Related Topics. https://doi.org/10.21236/ADA093824
40. Kim DH, Baddar WJ, Jang J et al (2017) Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. IEEE Trans Affect Comput 10:223–236
41. Guo Y, Zhang L, Hu Y et al (2016) Ms-celeb-1m: a dataset and benchmark for large-scale face recognition. In: Leibe B, Matas J, Sebe N et al (eds) Lecture notes in computer science, vol 9907. Springer, Cham. https://doi.org/10.1007/978-3-319-46487-9_6

42. Lucey P, Cohn JF, Kanade T et al (2010) The extended cohn-kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: 2010 IEEE computer society conference on computer vision and pattern recognition, 13–18 June, San Francisco, pp 94–101
43. Livingstone SR, Russo FA (2018) The ryerson audiovisual database of emotional speech and song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in North American English. PloS one 13:e0196391. https://doi.org/10.1371/journal.pone.0196391
44. Koelstra S, Muhl C, Soleymani M et al (2011) Deap: a database for emotion analysis; using physiological signals. IEEE Trans Affect Comput 3:18–31
45. Li W, Tan R, Xing Y et al (2022) A multimodal psychological, physiological and behavioural dataset for human emotions in driving tasks. Sci Data 9:481. https://doi.org/10.1038/s41597-022-01557-2
46. Zhang K, Zhang Z, Li Z et al (2016) Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett 23:1499–1503
47. Lawrence I, Lin K (1989) A concordance correlation coefficient to evaluate reproducibility. Biometrics 45:255–268
48. Deng S, Lv Z, Galván E et al (2023) Evolutionary neural architecture search for facial expression recognition. IEEE Trans Emerg Top Comput Intell 7(5):1405–1419
49. Rayhan Ahmed Md, Islam S, Muzahidul Islam AKM et al (2023) An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition. Expert Syst Appl 218:119633. https://doi.org/10.1016/j.eswa.2023.119633
50. Tang J, Ma Z, Gan K et al (2024) Hierarchical multimodal-fusion of physiological signals for emotion recognition with scenario adaption and contrastive alignment. Inf Fus 103:102129. https://doi.org/10.1016/j.inffus.2023.102129
51. Li W, Zeng G, Zhang J et al (2021) CogEmoNet: a cognitive-feature-augmented driver emotion recognition model for smart cockpit. IEEE Trans Comput Soc Syst 9(3):667–678

Ying-Zhang Wu received a B.S. degree in mechanical engineering from Chongqing University, Chongqing, China, in 2017. He is working toward a Ph.D. with the Advanced Manufacturing and Information Technology Laboratory, College of Mechanical and Vehicle Engineering, Chongqing University, Chongqing, China. His research interests include intelligent vehicles, intelligent cockpits, driver emotion detection, driving fatigue, human-machine interaction, and brain-computer interface.

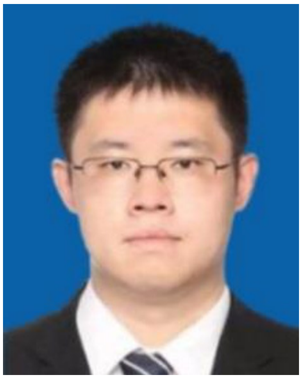
Wen-Bo Li received a B.S., M.Sc., and Ph.D. in automotive engineering from Chongqing University, Chongqing, China, in 2014, 2017, and 2021, respectively. From 2018 to 2020, he was a Visiting Ph.D. Student with the Department of Mechanical and Mechatronics Engineering, University of Waterloo, Waterloo, ON, Canada. From 2021 to 2023, he was a Postdoctoral Research Fellow at the School of Vehicle and Mobility at Tsinghua University, Beijing, China. He is an associate professor at the College of Mechanical and Vehicle Engineering, Chongqing University, Chongqing, China. His research interests include intelligent vehicles, intelligent cockpits, human emotion and cognition, driver emotion detection and regulation, human-machine interaction, affective computing, and brain–computer interface.

Yu-Jing Liu received her B.S. degree in industrial design from the College of Mechanical Engineering, Chongqing University, China, in 2020. She is working toward a Ph.D. with the Advanced Manufacturing and Information Technology Laboratory, College of Mechanical and Vehicle Engineering, Chongqing University, Chongqing, China. Her research interests include human-vehicle interaction design for intelligent cockpit, user experience, and car seat comfort.

Guan-Zhong Zeng received a B.S. and M.Sc. degree in automotive engineering from Chongqing University, Chongqing, China, in 2018 and 2021. He is an artificial intelligence engineer at Hikvision Research Institute, Hangzhou, China. His research interests include gaze estimation, domain generalization, and domain adaptation.[Inline Image Removed]Cheng-Mou Li received a B.S. degree in mechanical engineering from the College of Mechanical Engineering, Chongqing University, China, in 2020. He is working toward a Ph.D. with the Advanced Manufacturing and Information Technology Laboratory, College of Mechanical and Vehicle Engineering, Chongqing University, Chongqing, China. His research interests include intelligent transportation systems, human-computer interaction, driver distraction detection, and brain-computer interface.

Cheng-Mou Li received a B.S. degree in mechanical engineering from the College of Mechanical Engineering, Chongqing University, China, in 2020. He is working toward a Ph.D. with the Advanced Manufacturing and Information Technology Laboratory, College of Mechanical and Vehicle Engineering, Chongqing University, Chongqing, China. His research interests include intelligent transportation systems, human-computer interaction, driver distraction detection, and brain-computer interface.

Hua-Min Jin received an M.Sc. degree in Industrial Engineering from Seoul National University, Seoul, Korea, in 2019. She is an intelligent cockpit researcher at the China Society of Automotive Engineers in Beijing, China. Her research interests include intelligent vehicles, cockpits, human factors, user experience, and human-computer interaction.



Shen Li received a Ph.D. from the University of Wisconsin–Madison, USA, in 2018. He is a Research Associate at Tsinghua University. His research interests include intelligent transportation systems (ITS), architecture design of CAVH system, vehicle infrastructure cooperative planning and decision method, traffic data mining based on cellular data, and traffic operations and management.

Gang Guo received a Ph.D. degree in mechanical engineering from Chongqing University, Chongqing, China, in 1994. He is a professor at the College of Mechanical and Vehicle Engineering, Chongqing University, Chongqing, China. He has authored and co-authored over 100 refereed journal and conference publications. His research interests include human-machine interaction, user experience, intelligent cockpits, intelligent vehicles, brain-computer interfaces, and intelligent manufacturing.