IEEE Transactions on Computational Social Systems, Available online 30 November 2021
DOI:10.1109/TCSS.2021.3127935

CogEmoNet: A Cognitive-Feature-Augmented Driver Emotion Recognition Model for Smart Cockpit

Wenbo Li†, Guanzhong Zeng†, Juncheng Zhang, Yan Xu, Yang Xing, Rui Zhou, Gang Guo*, Yu Shen, Dongpu Cao, and Fei-Yue Wang, *Fellow IEEE*

***Abstract*—Driver's emotion recognition is vital to improving driving safety, comfort, and acceptance of intelligent vehicles. This paper presents a cognitive-feature-augmented driver emotion detection method which is based on emotional cognitive process theory and deep networks. Different from the traditional methods, both the driver's facial expression and cognitive process characteristics (age, gender, and driving age) were used as the inputs of the proposed model. Convolutional techniques were adopted to construct the model for drivers emotion detection simultaneously considering the drivers facial expression and cognitive process characteristics. A driver's emotion data collection was carried out to validate the performance of the proposed method. The collected dataset consists of 40 drivers' frontal facial videos, their cognitive process characteristics and self-reported assessments on driver emotions. Another two deep networks were also used to compare recognition performance. The results prove that proposed method can achieve well detection results for different databases on the discrete emotion model and dimensional emotion model, respectively.**

***Index Terms*—Driver emotion, Human-machine interaction, Affective computing, Smart cockpit, Facial expression**

I. Introduction

With the advancement of artificial intelligence and computing systems, intelligent vehicles applications have been growing rapidly worldwide. Together with the development of communication technologies, extensively emerging technologies have been developed to connect with vehicles, pedestrians, infrastructures and clouds in the transportation network [1], [2]. Thus, intelligent vehicles have become multi-intelligence mobile terminal that carries rich functions and services [3]. This multi-intelligence mobile terminal will bring about tremendous changes in the interaction system of automotive smart cockpits. The automotive smart cockpit is an intelligent service system equipped with intelligent and connected in-vehicle products or technologies with the ability of insight, understanding, and meeting user needs in the application scenarios to achieve safe, efficient, comfortable, and pleasant human-machine interaction (HMI) experience. The development of the smart cockpit will expand and deepen the scope of HMI between humans and vehicles, resulting in new human-vehicle interaction problems that challenge safety, comfort, and driver's acceptance [4]. Among the problems, emotion-aware human-vehicle interactions are urgently needed to be addressed for improvement [5].

Driver's emotion recognition and regulation are the main topics for emotion-aware human-vehicle interactions [6]. Previous studies have shown that emotion recognition and regulation systems can be used to understand and regulate the emotional state of the driver to enhance the safety, comfort, and acceptance of driving [6], [7]. The emergence of intelligent cockpit HMI technology brings new thoughts to solve the emotional disorder problems of the drivers. Drivers' emotions can be recognized and regulated in various ways by the emotion-aware HMI of the smart cockpit, based on which the safety and comfort of driving can be improved. As the first step for the development of the emotion-aware HMI system, precise driver emotion recognition is of great significance to realize the above improvement [4].

Driver emotion recognition is usually carried out by analyzing the driver's emotional expression. The emotions of humans can be expressed in forms including behavioral expressions and physiological changes. Up to present, various behavioral measurements [8], [9], physiological signal measurements [10], [11], or self-reported scales [4], [12] have been applied to driver emotions recognition. Considering the significant impact of the interference and intrusion on the emotional expression and real emotion experience, the application of non-invasive, non-contact, continuous measurement methods during the study of driver emotions is essential [4], [13]. Therefore, in this study, the facial expressions of the driver are adopted as the main information to recognize the drivers emotion.

At the same time, driving is a complicated cognitive process

Wenbo Li, Guanzhong Zeng, Juncheng Zhang and Gang Guo are with the College of Mechanical and Vehicle Engineering, Chongqing University, Chongqing, 400044, China (e-mail: liwenbocqu@foxmail.com, guanzhong@cqu.edu.cn, zhangjuncheng@cqu.edu.cnguogang@cqu.edu.cn).

Yan Xu is with the Department of Mechanical Engineering, University of Science and Technology Beijing, 100083, Beijing, China. (e-mail: b20160225@xs.ustb.edu.cn).

Yang Xing is with the Department of Aerospace, Transport, and Manufacturing, Cranfield University, Cranfield, MK43 0AL, UK. (e-mail: Yang.X@cranfield.ac.uk).

Rui Zhou is with the Department of Research and Development, Waytous Inc., Shenzhen, China. (e-mail: rui.zhou@waytous.com).

Yu Shen is with the School of Artifitical Intelligence, University of Chinese Academy of Science, Beijing 100049, China (e-mail: shenyu2015@ia.ac.cn).

Dongpu Cao is with the Department of Mechanical and Mechatronics Engineering, University of Waterloo, Waterloo, ON, N2L 3G1, Canada. (e-mail: dongpu.cao@uwaterloo.ca).

Fei-Yue Wang is with the State Key Laboratory of Management and Control for Complex System, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: feiyue@ieee.org).

† These authors contributed equally

*Corresponding author

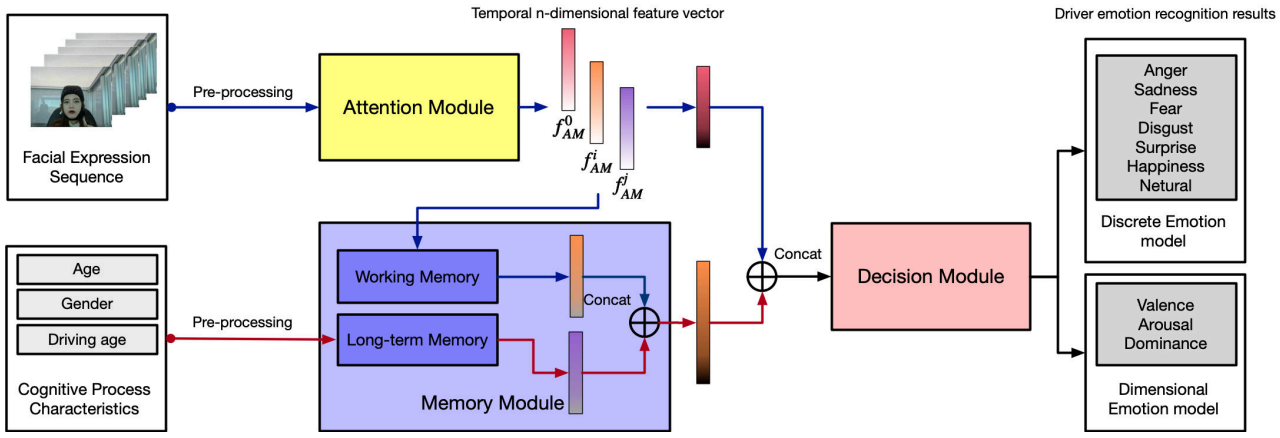


Fig. 1: Illustration of the framework of CogEmoNet model

requiring dynamical responses to the driving task which occupies a great quantity of cognition and requires the cognitive appraisal to trigger emotional responses [7], [14], [15], [16]. The driver's facial expression may be suppressed due to the influence of the driving task. In addition, the impact of the cognitive process is closely related to the driver's age, gender, and driving experience [4]. Based on this, the facial expressions and cognitive process of the driver should be considered during the study of the drivers emotion recognition under the dynamic driving scenario.

To resolve the limitation introduced above, this study proposed a model based on facial expression and cognitive process for driver emotion recognition of smart cockpit. Fig. 1 shows the structure of the cognitive-feature-augmented driver emotion recognition network (CogEmoNet) model.

The main contributions of the study can be concluded as follows:

- A cognitive-feature-augmented recognition model of driver emotions named CogEmoNet is proposed. This model is proposed and implemented by combining emotion generation process theory and deep learning algorithms. CogEmoNet recognizes driver emotions by simultaneously considering the drivers facial expression and cognitive process characteristics (age, gender, and driving age).
- This study conducted driver's emotion data collection. The collected driver's emotional facial expression (DEFE+) dataset is composed of frontal facial videos, cognitive process characteristics, and subjective ratings on emotions of 40 drivers. The cognitive process characteristics include the information of age, driving age, and gender. The subjective ratings include the information of valence, arousal, dominance, and seven emotion categories.
- The effectiveness of the proposed CogEmoNet driver emotion recognition framework was evaluated on DEFE+ dataset. It was also evaluated using leave-one-out cross-validation on other publicly available and widely used CK+ and DEAP databases. Furthermore, a comparison between the CogEmoNet and state-of-the-art models is performed to prove that the CogEmoNet performs significantly well in driver emotions recognition.

The structure of this paper is as follows: related works about emotion recognition are summarized in Section II. The proposed CogEmoNet is introduced in detail in Section III. Section IV introduces the process of data collection of DEFE+. The experiment results of CogEmoNet are analyzed in Section V. The conclusion are in Section VI.

II. Related Research

A. Emotion Classification

To describe human emotions, psychological researchers have proposed discrete emotion theory and dimensional emotion theory to classify emotions [17]. At present, the most acknowledged discrete emotion model is the basic emotion model proposed by Ekman [18]. Other emotions were regarded as combinations of these basic emotions. Dimension emotion theory points out that psychological dimensions including valence, arousal, dominance. can be combined to accurately express human emotion. Specifically, whether a person feels positive or negative, whether a person feels bored or excited, and whether a person feels submissive or empowered are measured by the dimension of valence, arousal, and dominance respectively [19], [20].

The widely used discrete emotion method can intuitively reflect the emotions, but only several limited emotions are included. The dimensional emotion method has the advantage of high practicability and context-sensitive, but is less intuitive and requires a more complex process of labeling the data, [17]. In this paper, with the help of the differential emotion scale (DES) [21] and the self-assessment model (SAM) [22], the discrete emotion method and dimensional emotion method are combined.

B. Emotional Cognitive Process

According to the cognitive theory [29], an emotional response begins with an cognitive appraisal of the personal significance of a situation. This cognitive appraisal further leads to the emotional response, including subjective experience, physiological change and behaviour response [18], [30]. Therefore, the cognitive process and emotional expression are

TABLE I: A summary of representative studies of deep-learning-based facial expression detection methods

Author	Emotions	Network	Datasets
Liang et al. [23]	7 emotions	CNN, LSTM	CK+, Oulu-CASIA, MMI
Li et al. [24]	7 emotions	CNN	CK+, JAFFE
Ivan, Gogi et al. [25]	7 emotions	LBF-NN	CK+, MMI, JAFFE, SFEW
Wang et al. [26]	8 emotions	CNN	FERPlusAffectNet, SFEW, RAF-DB
Wang et al. [27]	7 emotions	SVM	Jaffe, CK+, BU-3DFE
Liu et al. [28]	7 emotions	CNN, LSTM	CK+, Oulu-CASIA, MMI, AffectNet, AFEW

both essential for driver's emotion recognition. The human information processing stage model [31], as shown in Fig. 2, is used to analyze the psychological processes while the subjects are interacting with the system and performing tasks. A series of processing stages or mental operations, which typically (but not always) characterize the information flow as humans perform tasks, can be described by the model.

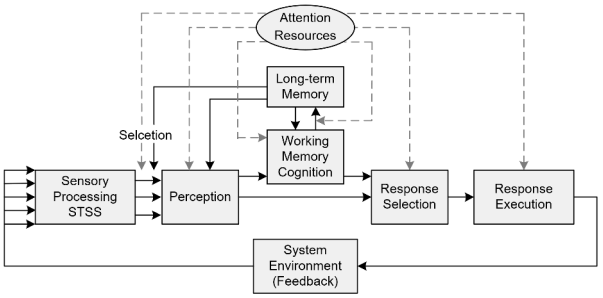


Fig. 2: The model of human information processing stages [31]

Consider as an example of a driver who is afraid of a traffic accident while driving. As shown on the left side of Fig. 2, events in the environment are processed by the senses, namely sight, sound, etc., and may be stored in the short-term sensory store (STSS) briefly. Perception involves determining the meaning of the sensory signals or events, and such meaning comes from past experience (cautious when encountering traffic accidents), stored in the long-term memory, including facts, images, and the running pattern of the world. There are two ways in which the information may be processed after perception. In the first way, the response is executed after the stages of response selection and execution, which is related to the muscles and the way the brain controls them. Compared with the first way, in which reactions may not always get triggered immediately by perception and the understanding of the situation, in the second way, the state of event is temporarily retained by the driver, using working memory, and more information, namely approaching vehicles or possible police cars, etc. is obtained as the driver is scanning the road ahead at the same time. Note that attention is an important part of most information processing.

We describe the process after perception and before response execution as the "cognitive process". Combining the human information processing stage model and emotion-related theories, the process of human emotion generation can be divided into four stages, namely perception, cognition, expression, and feeling. Perception includes events that occur in the current environment that humans perceive. It is obtained by our sensory organs, including visual information, auditory information, tactile information, etc., and all the perceived content is used as cognitive input. The cognitive process mainly includes four components: attention resources, long-term memory, working memory cognition, and response selection, which can be summarized as three important stages of attention, memory, and decision. Attention can filter some unimportant information, so that our limited cognitive resources can be used to process main tasks, and then combined with long-term memory and working memory to select and execute the response. Expression includes facial expressions, behavioral actions, speech, and physiological reactions, and finally, produce corresponding emotional feelings. This study employed deep networks to simulate the process of human cognitive information processing, so that intelligent vehicles can recognize driver emotions in the way of the human cognitive processing, to build a computational model for driver emotion recognition.

C. Facial Expressions-based Emotion Recognition

Convolutional neural network (CNN) and recurrent neural network (RNN) have been applied to facial emotion recognition in recent years due to the advancement of computing power and deep learning algorithms. CNN is suitable for parallel computing, which can directly extract deep features from the input image, and directly carry out the recognition task without manual feature construction, and no longer rely on expert experience and data processing techniques. Most methods based on deep learning use CNN to detect action unit (AU) directly. For example, Breuer and Kimmel [32] verified the emotion detection ability of various facial emotion recognition networks through CNN visualization technology. Jung et al. [33] used a dual-stream network to extract temporal appearance features and temporal geometry features to improve facial expression recognition ability.

In addition to CNN being directly used for facial feature extraction, many methods begin to combined CNN and long short-term memory (LSTM) to recognize facial emotions in video sequences since RNN or LSTM is more suitable for constructing temporal features. The hybrid CNN-LSTM model is flexible because LSTM supports input or output of fixed and unfixed length. For example, Kim et al. [34] used the beginning, peak, and end of facial expression to represent the expression state by using CNN to extract spatial features and LSTM to learn temporal features and combining with spatial and temporal features to recognize facial emotions. Chu et al. [35] predicted 12 AUs of each video frame through the hybrid CNN-LSTM model. Hasani and Mahoor [36] proposed a 3D perception ResNet model to emphasize different facial regions.

Liang et al. [23] achieved excellent performance through joint learning of facial expressions and temporal dynamics. Li et al. [24] proposed a new data augmentation method for the small amount of emotional facial expression dataset, using face cropping and rotation to improve network accuracy. Ivan, Gogi et al. [25] optimized facial expression recognition by concatenating sparse facial expression binary features; Wang and Peng et al. [26] proposed a new region attention network and region bias loss to improve the recognition robustness of facial occlusion and pose changes; Wang and Li et al. [27] proposed a new facial expression representation method to reflect the characteristics of local expression, texture, appearance, and shape; Liu et al. [28] proposed a metric learning framework with a siamese cascaded structure. Some representative works among various methods based on CNN or hybrid CNN-LSTM are shown in Table I. In this paper, to simplify the realization of the theoretical model, we used CNN as the component to build CogEmoNet. Besides, the CNN-LSTM method was employed to compare the performance of CogEmoNet.

III. Model Description

The CogEmoNet model is a driver emotion recognition model based on facial expression and cognitive process, it further modeled the feature extraction stage. The model takes the temporal driver's facial expression images and cognitive process characteristics (age, gender, and driving age) as input and finally outputs the recognition results for the discrete emotion model (anger, sadness, fear, disgust, surprise, happiness, neutral) and dimensional emotion model (arousal, valence, dominance) to recognize the driver's emotions.

A. Overall framework of the CogEmoNet model

The computational model for driver emotion recognition based on the facial expression and emotional cognitive process proposed in this paper mainly includes three stages of attention, memory, and decision. Correspondingly, CogEmoNet is divided into three modules: attention, memory, and decision, as shown in Fig. 1. The first module is the attention module, which realizes the emphasis and filtering of input information. The module's input is temporal facial expression images, using CNN as the basic feature extractor and obtaining more discriminative features by adding spatial and channel attention. The second module is the memory module. In order to generate memory information representing the driver's experience and the current scene to help the final decision, the memory module separately constructs the driver's long-term memory and working memory, composed of multi-layer perceptrons (MLP) and exponentially weighted moving average operations (EWMA). Driver's cognitive process characteristics (age, gender, and driving experience) are used to generate long-term memory features, an EWMA of the temporal facial features output by the attention module to generate working memory features. Finally, the third module is the decision module, which uses the fully connected layer (FC) to combine the features extracted by the attention module and memory module to detect the driver's emotions.

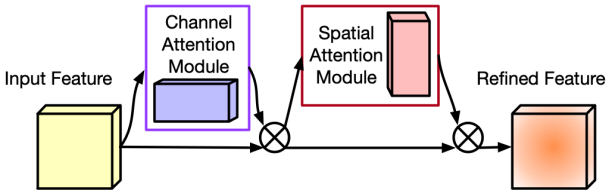


Fig. 3: The basic block of attention module: CBAM_ResBlock [37]

One of the inputs of CogEmoNet is processed temporal image frames. Face landmarks on the original image are processed, then face alignment operation is conducted to crop and resize the image to target size so that the center of the driver's eyes and the mouth is kept at a fixed position the image. Next, the processed images are input into the attention module, which is a CNN with channel and spatial attention, without the FC and classification layer. Finally, the deep features extracted by the attention module and the driver's cognitive process characteristics are sent to the memory module together. The memory module is divided into long-term and short-term memory, simulated by MLP and EWMA, respectively. The input of MLP is the driver's cognitive process characteristics, which are used to construct the driver's specific long-term memory characteristics. The input of EWMA is the driver's facial features extracted from all frame images, which reflects the changing trend of the driver's facial features and is used to represent the driver's working memory features. Notably, besides the output features of MLP and EWMA, the driver's facial features output by the attention module are also added to the decision module, a FC with a classification/regression function to detect the driver's emotions on the input video.

B. Attention Module

This study used the widely used ResNet [38] to extract facial features and spatial channel attention convolutional block attention module (CBAM) [37] to weight the features according to their importance (as shown in Fig. 3), thereby improving the models representation ability. The study integrated the CBAM module into ResNet34, which was removed the decision layer, as the calculation model of the attention module. The calculation process is as follows:

$$f_{AM}^{i} = F_{AM}(f_{input}^{i}), i \in [0, k-1] \tag{1}$$

Which, F_{AM} is the function of attention module, $f_{input}^{i}, i \in [0, k-1]$ is the i-th input image, k is the number of video frames, f_{AM}^{i} is the facial feature vector of the driver in the i-th frame.

The facial feature vector extracted by the attention module will be input to the memory module and averaged on the channel and then input to the decision module.

C. Memory Module

The memory module is composed of two parts, as shown in Fig. 4. Part of it is implemented by a MLP, which inputs the drivers cognitive characteristics, including age, gender,

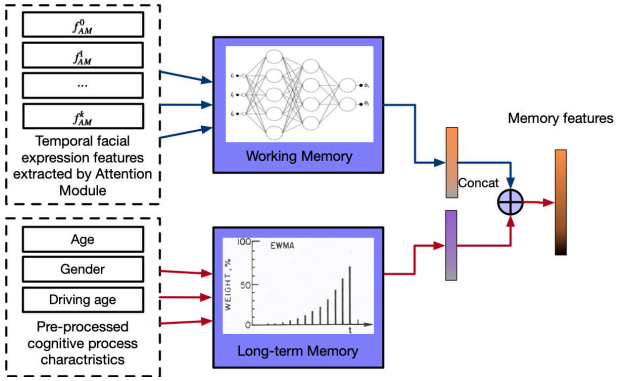


Fig. 4: Memory Module

and driving age, and outputs driver-specific cognitive characteristics, which represents the drivers long-term memory characteristics; the other part of it is realized by EWMA. The input is the temporal facial features output by the attention module. After EWMA processing is performed channel by channel according to the time sequence, the output represents the driver's working memory characteristics of the current video sequence. Finally, the memory module receives two independent inputs and generates two independent outputs, and finally combines the long-term memory features and working memory features as the driver's complete memory features. The calculation process is as follows:

$$f_{MLP} = F_{MLP}(f_{gender}, f_{age}, f_{driving\_age}) \tag{2}$$

$$f_{EWMA} = F_{EWMA}(f_{AM}^{0}, ..., f_{AM}^{k}) \tag{3}$$

$$f_{MM} = concat(f_{MLP}, f_{EWMA}) \tag{4}$$

Which f_{gender}, f_{age}, $f_{driving\_age}$ represent the drivers gender, age and driving age, respectively, F_{MLP} is the function of the multilayer perceptron, F_{EWMA} is the exponentially weighted movement average functions, f_{MLP}, f_{EWMA}, f_{MM} correspond to the drivers long-term memory characteristics, working memory characteristics, and completed memory characteristics, respectively.

D. Decision Module

The decision module is implemented by the FC as a "classifier/regressor" in the entire model. According to the human information processing stage model, the input of response selection includes attention and the output of memory cognition. Therefore, the input of the decision module is the combination of the output of the attention module and the memory module, as shown in Fig. 1.

E. Loss Function

Different loss functions are used for different emotion recognition tasks. Cross entropy (CE) [39], F1 [40], mean square error (MSE) [41], and consistency correlation coefficient (CCC) [34] loss function are applied to optimize accuracy (Acc), F1 score, MSE, and CCC respectively. The corresponding formulas are listed below.

$$L_{CE\ loss} = -\frac{1}{N}\sum_{i=1}^{N} log(\frac{e^{h_{yi}}}{\sum_{j=1}^{C} e^{h_j}}) \tag{5}$$

$$L_{F_1\ loss} = 1 - 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

$$L_{MSE\ loss} = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{M}(I_t^i)^2 \tag{7}$$

$$L_{CCC\ loss} = 1 - \frac{2S_c}{S^2 + \hat{S}^2 + (\bar{y} - \bar{\hat{y}})^2} \tag{8}$$

Where x_i is the input feature of the $i-th$ sample in the final classification layer, $y_i \in \{1, 2, ..., C\}$ and $\hat{y_i} \in \{1, 2, ..., C\}$ are corresponding true label and predict label of the $i-th$ sample respectively, $\bar{y}$ and $\bar{\hat{y}}$ is their corresponding average. S, $\hat{S}$, and S_c is the variance and covariance of y_i and $\hat{y_i}$. $h = (h_1, h_2, ..., h_C)^T$ is output of the network, namely the recognition of the $i-th$ sample, C is the number of classes.

IV. Data Collection

To verify the effectiveness of CogEmoNet model, a dataset including drivers facial expressions and cognitive process characteristics needs to be collected. In this section, we collected the DEFE+.

A. Ethics Statement

The video-audio clips' content shown to the participants and the whole experimental procedure were approved by the Ethics Committee of Chongqing University Cancer Hospital, China. Participants and data from participants were treated according to the Declaration of Helsinki.

B. Stimulus Selection

The emotions of the driver need to be induced by the appropriate stimulus in order to collect the emotion data. Video-audio clips have been proved to be reliable to elicit the emotions of the driver [3], [4]. In this paper, forty-two clips were manually selected, referring to the methods of previous research [4]. Participants were recruited to rate these clips subjectively, based on which seven clips were selected.

1) Participants: We recruited fifty participants with driving experience for more than one year from Chongqing University, 9 of whom are female and the rest are male. All the participants have valid driver's licenses. The age of participants ranges from 21 to 32. Their driving age ranges from 1 to 10 years. The average age and average driving age are 25.3 and 3.5 respectively. The standard deviations of age and driving age are 2.6 and 2.2 respectively. Agreement to participate in the study was signed by all fifty participants.

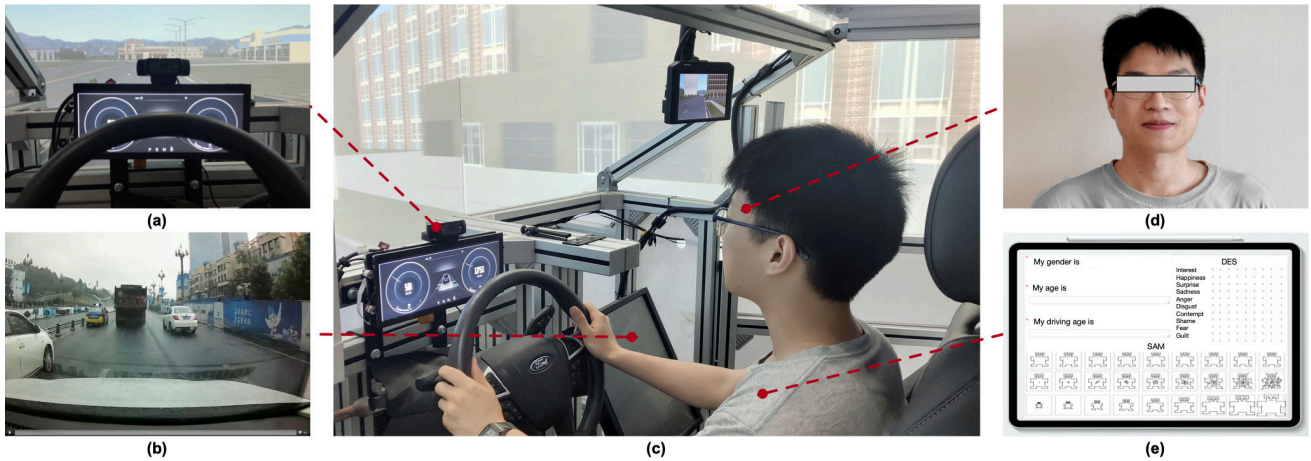


Fig. 5: Overall data collection setup. (a) visual face camera, (b) video-audio stimulus display, (c) data collection setup, (d) driver facial expression recording, (e) driver cognitive process characteristics data collecting and his/her self-reported emotion

2) Materials and Procedure: Subjective annotation was achieved via SAM and DES. The level of arousal, valence, and dominance was indicated with non-verbal graphical representations by SAM. DES was used to assess personal emotions [42]. The effectiveness of SAM was proved in a previous study [22]. In this paper, 9-point scales (1 = not at all, 9 = extremely) [21] of SAM [22] and DES were adopted to evaluate the intensity of each self-reported emotional dimension.

The clips were rated via a subjective emotion assessment experiment. A set of instructions was provided to each of the participants to explain the definition of SAM and DES before the experiment. Two questionnaires were finished by each of the participants according to their true feelings after watching each clip. Forty-two clips were randomly displayed. For each clip, 50 assessments were collected.

3) Selection Results: Both results of SAM and DES were used to pick out the most effective seven clips. In the process of analyzing the data of SAM, in order to elicit the emotions with the maximum strength, the average score of each clip was calculated, and the clip with the highest average score and small variation was selected. The normalized score of valence, arousal, and dominance of each clip was clustered with the K-means method, thus the emotion clusters were identified based on SAM data [3], [4]. Besides, the hit rate, intensity value, and success index was defined according to DES result to select the clips which were effective to induce the emotions of the driver [3], [4]. The selection results of SAM and DES were analyzed to be consistent. Seven clips were selected. Table II shows the contents of the clips.

TABLE II: The content and duration of selected clips for driver emotion induction

Target emotion	Content	Duration (sec)
Anger	The driver is driving on the road and intentionally jammed by other cars	30
Sadness	The driver heard the latest report of the earthquake broadcast on the radio while driving	63
Fear	Serious traffic accidents on the road	59
Disgust	The driver noticed that the rear passenger's slippers put his feet on the co-pilot's position	57
Surprise	The traffic police investigated a van with more than 50 people in it	94
Happiness	A collection of various modified vehicles driving on the road	79
Neutral	The driver drives on wide city roads with nothing happened	48

C. Facial Expression and Cognitive Data Collection

1) Participants: We recruited forty Chinese participants with driving experience for more than one year, 9 of whom are female and the rest are male. All the participants have valid driver's licenses. The age of participants ranges from 19 to 55. Their driving age ranges from 1 to 32 years. The average age and average driving age are 28.03 and 5.58 respectively. The standard deviations of age and driving age are 9.24 and 6.02 respectively. The experiment was carried out in Chongqing. All participants had normal or corrected to normal vision and normal hearing ability.

2) Apparatus and Driving Scenarios: The driving experiments were implemented in a driving simulator with illumination-controlled (RDS2000). A screen was adopted, in which the clips were displayed. The resolution ratio of the screen is $1,280 \times 1,024$ and the refresh rate is 60Hz. The video data was collected with a Pro Webcam C920 (Logitech), of which the resolution is $1,920 \times 1,080$ pixels as the visual face camera, the frame rate of which is 30 fps. The self-reported emotion and the cognitive process characteristics data, namely age, gender, and driving experience of the participants was collected with an iPad (Apple). The overall setup of data collection is shown in Fig. 5.

Two driving scenarios on highways, as shown in Fig. 6 for practice and formal experiment respectively, were realized in the simulator. The reason for setting these two scenarios is to minimize the impacts of complex driving conditions on the driver's emotion experience and their facial expressions [43]. Both scenarios were highways with four lanes, two for one direction and another two for the opposite direction. The length of the practice and formal experiment highway was 8km and 3km respectively. The participants were asked to drive in the right lane. To make the participants familiar with the

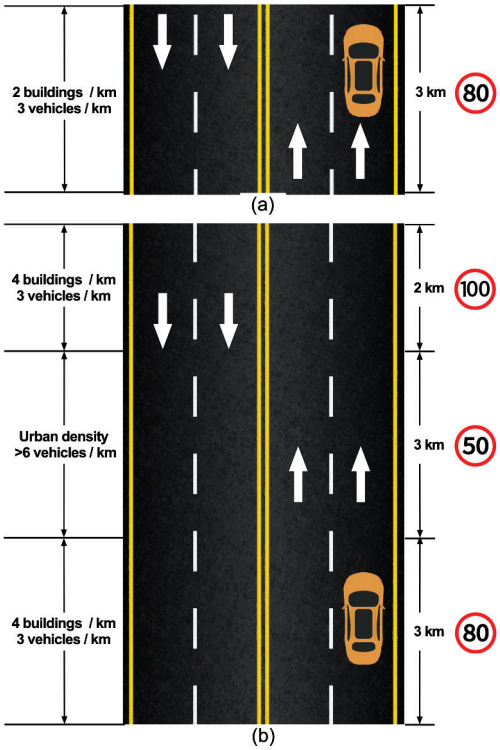


Fig. 6: Illustrations of driving scenarios: (a) emotion driving, (b) familiarization driving

equipment, they were asked to drive on the practice highway and controlling the speed at 80km/h, 50km/h, and 100km/h at different times. During the formal experiment, the participants were required to control the speed at 80km/h all the time until they finish the experiment.

3) Data Collection Protocol: Before the experiment, a set of instructions was taught to each participant to explain the experimental protocol and details of the self-reported emotion scale. After a ten-minute familiarization driving followed by a short break, seven emotion drivings, namely angry driving, sad driving, fear driving, disgust driving, surprise driving, happy driving and neutral driving were started in random order with a three-minute break between each emotional driving. The corresponding emotion of each emotional driving was induced with the selected clip at the beginning of the emotional driving. The self-evaluated emotion level was reported in the form of SAM and DES by the participants when each emotional driving was finished. The faces of the participants were continuously recorded by the camera during each emotional driving. After the entire experiment, each participant filled out a questionnaire to collect their cognitive process characteristics data (age, gender, and driving age). In sum, each participant drove seven times in the highway simulation scene with the data recording. Therefore, for 40 participants, 280 times driving were finished with data collection. Each participant took about 90 minutes to complete the entire process of data collection. The average time of the whole data collection section of one participant was 945s. Each participant took about 90 minutes to complete the entire process of data collection.

D. Target Emotion Induction Success Check

The DES of each participant was used as the ground truth to verify whether the target emotion was generated by the participant during the emotional driving. The self-reported emotion would be selected as the ground truth when it was not consistent with the target emotion.

It was shown in the results that for each of the emotional drivings, namely angry, sad, fear, disgust, surprise, happy, and neutral driving, 34, 38, 36, 25, 34, 36, and 37 participants were successfully induced into the target emotion, respectively. 240 participants were successfully induced in total. Notably, in DEFE+ dataset, we removed the facial expression and cognitive data that was not successfully induced.

V. Experimental Setup and Results

Based on the above models and datasets, a driver emotion recognition algorithm based on cognitive process theory and facial expressions can be realized. The performance advantages of CogEmoNet were verified on different datasets and different evaluation metrics.

A. Datasets Used

Since the facial expression datasets with video sequences are generally small, To improve the feature extraction ability and expression recognition performance of the attention module, the MS-Celeb-1M dataset [44] was used to pre-train the attention module. In addition, the CogEmoNet proposed in this paper can be used to identify discrete emotions and dimensional emotions, and the DEFE+ dataset collected in section IV covered the truth labels of discrete emotions and dimensional emotions, in order to verify the universality of CogEmoNet, The performance verification of CogEmoNet with different evaluation metrics was also carried out on the CK+ [45] dataset with discrete emotion labels and the DEAP [46] dataset with dimensional emotion labels. A sample of each dataset was shown in Fig. 7.

DEFE+: The DEFE+ dataset covered the facial expression in the driving scenario and the driver's cognitive process characteristics. According to six basic emotions and neutral emotion and arousal-valence-dominance dimensional emotions, 240 video sequences were successfully induced by 40 participants and labeled. The 15s facial expression video sequence after driving was edited as the most effective data [4]. Due to various postures, lighting, and occlusion (glasses), face detection and alignment in a driving scenario was challenging. The resolution of the image was 640×480.

CK+: Almost 600 video sequences of 123 participants were included in CK+. The ages of participants, most of whom were women, range from 18 to 30. 327 video sequences of 118 people were labeled with seven emotion labels. The emotion classification results of the discrete models were compared using CK+ as one of the datasets. The resolution of the image was 640×480 and 640×490.

DEAP: The physiological signals of 32 participants (peripheral physiological information, electroencephalogram, and frontal facial data of 22 participants) were included in DEAP. After watching each of forty stimuli chooses to induce a

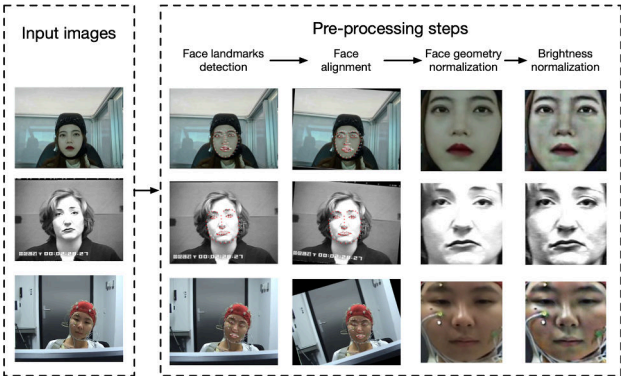


Fig. 7: Sample images of each dataset and data pre-processing setps

certain emotion respectively, the real emotions of each participant were evaluated by themselves from arousal, dominance, valence, liking, and familiarity. The score of the first three dimensions range from 1 to 9, and the other two dimensions range from 1 to 5. The bigger the number, the stronger the emotion. During the experiment, the facial videos of twenty-two participants were recorded, which were adopted to compare the results of dimensional emotion models. There were a total of 880 facial video sequences. The resolution of the image was 720 $\times$ 576.

B. Data Pre-processing

1) Facial Expression Sequence Data: All the experimental datasets compared in this paper contain face video sequences. Due to different image collecting methods, the input image had complex illumination conditions and large head postures. In addition, the distance, focal length, etc. made the size and position of the face in the entire image uncertain. In order to ensure the consistency of face size, position, and image quality, a series of pre-processing operations were applied to the input images.

Data pre-processing mainly included landmarks detection, face alignment, geometry normalization, and brightness normalization of the face image, as shown in Fig. 8. Face alignment was applied to get the face image with the correct face position. Geometric normalization was used to obtain standardized face images with the same size and facial area. Brightness normalization improved the quality of the image and made the image more suitable for human observation and computer processing and recognition.

First, multi-task cascade convolution network (MTCNN) was used to detect the 68 landmarks of the face [47]. The angle of two lines, namely the central line of two eyes and horizontal line, is used to rotate the image to align face; then, based on the distance a between two eyes' center, and the distance b between the mouth center and the center of two eyes, crops the driver's face image to width 2a and height 2b, and then resizes to 112$\times$112 pixels, as shown in Fig. 8. Geometric normalization made the same facial landmarks approximately located at the same region. At the same time, this process discarded the background details and facial regions such as ears and forehead that were not related to facial expressions [48], because these regions did not represent the specific information of facial expressions [49]. To reduce the change of the image signal caused by the change of illumination, the brightness of the cropped face image was normalized.

Previous studies confirmed that facial expressions usually last for 0.5 to 4 seconds [50], and we may sample at least two frames of images per second to capture changes in facial expressions. Therefore, for DEFE+ dataset (each video has a duration of 15 seconds), we collected a total of 30 frames. Meanwhile, to keep the model input consistent, the CK+ and DEAP datasets were also sampled to 30 frames. Notably, because the CK+ dataset provided video sequences with an indefinite number of frames, for video sequences with less than 30 frames, we used up-sampling to 30 frames, and for video sequences with more than 30 frames, we used down-sampling to 30 frames.

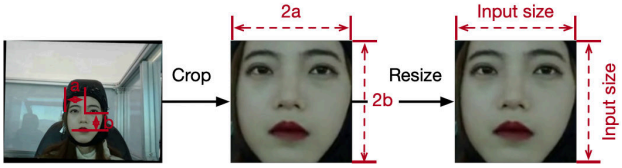


Fig. 8: Facial geometric normalization

2) Cognitive Characteristics Data: The pre-processing of driving cognitive characteristics data in DEFE+ was only used to generate the driver's long-term memory features. Among them, gender was binarized, and driving age and age were standardized with 0 mean and 1 variance.

For the CK+ and DEAP datasets without driver cognitive characteristics, we manually marked the gender of the current sample. The age and driving age were set to 0 and no additional data pre-processing was required.

C. Experiment Details and Evaluation Metrics

1) Experiment details: This model was implemented using the open-source platform PyTorch. The model was trained and tested on a server equipped with NVIDIA Tesla V100 GPU. CogEmoNet used Stochastic Gradient Descent (SGD) optimizer with Nesterov momentum of 0.9, the batch size was 64, and the learning rate decay strategy was stochastic gradient descent with warm restarts (SGDR) [51]. For CK+, DEAP, and DEFE+ datasets, the initial learning rate was 0.05, 0.005, and 0.0005 respectively. The model training was executed for 30, 60, and 30 epochs respectively with the warm restart of 5 epochs.

The dataset used in this paper were constructed based on 10-fold person-independence cross-validation to ensure that the tasks were independent of each other. Ten subsets were constructed as in several previous works [35]. The final experimental result was the average result of 10-fold cross-validation.

2) Evaluation metrics: The Basic discrete emotion model includes seven emotions (anger, sadness, happiness, neutral, fear, disgust, surprise), therefore, the recognition of discrete emotions was a multi-classification task, and Acc was the most

commonly used evaluation index in classification tasks. To deal with the category imbalance in CK+ and DEFE+, F1 score was added as a supplementary evaluation index of discrete emotion.

The dimensional emotion model includes three dimensions: arousal, valence and dominance, and the recognition of the dimensional emotion model was a regression task. The MSE was the most commonly used evaluation index in regression tasks. Since the CCC was also often employed to assess the effectiveness of emotion recognition, recognition results of dimensional emotion on DEAP and DEFE+ will were analyzed from the aspects of MSE and CCC. MSE was used to measure the overall mean deviation between the true value θ and its estimate $\hat{\theta}$. Notably, the smaller the MSE, the better of the model performance. CCC was a commonly used metrics in dimensional emotion recognition, and it was used to measure the consistency between the real emotion and the predicted emotion. The value of CCC ranges from -1(completely inconsistent) to 1(completely consistent). The formula of CCC is given below.

$$\hat{\rho_c} = \frac{2S_c}{S^2 + \hat{S}^2 + (\bar{y} - \bar{\hat{y}})^2} \tag{9}$$

D. Experimental Result

All methods in this paper used a fixed number of expression frames. The model structure of the baseline method was CBAM_ResNet34, which averages the recognition results of all expression frames in the video sequence as the recognition result of the video. Besides, the hybrid CNN-LSTM model was the commonly used method for facial expression detection, so the hybrid CNN-LSTM model was also used as the comparison method of the models proposed in this paper. The hybrid CNN-LSTM model employed CBAM_ResNet34 to extract spatial features of facial expressions, and LSTM to extract temporal features from the spatial features to predict emotions.

TABLE III: Recognition results in comparison of discrete emotion model on CK+ and DEFE+ datasets

Models	CK+		DEFE+	
	Acc	F1 score	Acc	F1 score
Baseline	0.851	0.819	0.295	0.256
CNN-LSTM	0.882	0.837	0.324	0.298
CogEmoNet (Our)	**0.907**	**0.882**	**0.351**	**0.327**

1) Discrete Emotion Model: The experimental results of discrete emotion on CK+ and DEFE+ were shown in Table. III. The higher the Acc and F1 score, the better the performance. We present the Acc and F1 score of discrete emotion recognition in Fig. 9(a) and Fig. 9(b) respectively. Table. III demonstrates that the results of CogEmoNet are better than baseline and CNN-LSTM methods, and performs best on CK+ and DEFE+. On CK+, the CogEmoNet Acc (90.7%) is 5.6% and 2.5% higher than the baseline (85.1%) and hybrid CNN-LSTM (88.2%) respectively. The CogEmoNet F1 score (88.2%) is 6.3% and 4.5% higher than the baseline (81.9%) and hybrid CNN-LSTM (83.7%) respectively. On DEFE+, the CogEmoNet Acc (35.1%) is 5.6% and 2.7%

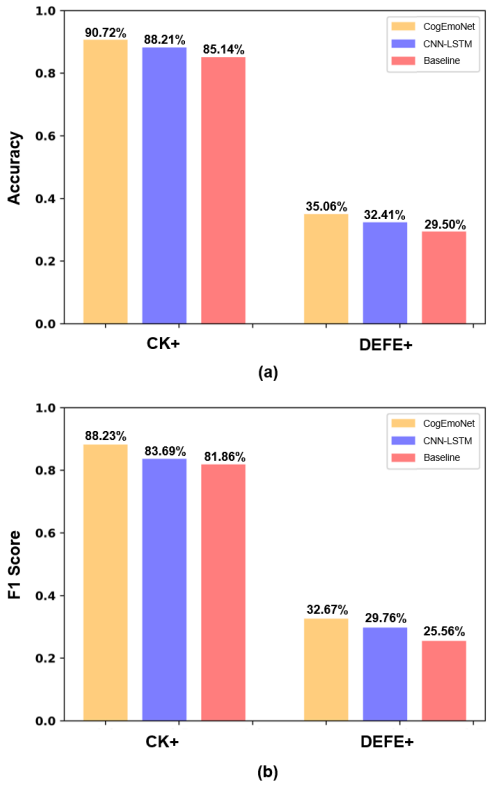


Fig. 9: Acc and F1 score of discrete emotion model

higher than the baseline (29.5%) and hybrid CNN-LSTM (32.4%) respectively. The CogEmoNet F1 score (32.7%) is 7.1% and 2.9% higher than the baseline (25.6%) and hybrid CNN-LSTM (29.8%) respectively. The effectiveness of our proposed CogEmoNet in discrete emotion recognition was proved.

TABLE IV: Recognition results in comparison of dimensional emotion model on DEAP and DEFE+ datasets

Models	DEAP		DEFE+	
	MSE	CCC	MSE	CCC
Baseline	8.447	0.117	9.902	0.155
CNN-LSTM	4.739	0.140	5.802	0.204
CogEmoNet (Our)	**3.654**	**0.181**	**4.220**	**0.221**

2) Dimensional Emotion Model: The experimental results of dimensional emotion on DEAP and DEFE+ were shown in Table. IV. The lower the MSE and the higher the CCC, the better the performance. We present the MSE and CCC of dimensional emotion recognition in Fig. 10(a) and Fig. 10(b) respectively. Table. IV shows that the results of CogEmoNet are better than baseline and CNN-LSTM methods, and performs best on DEAP and DEFE+. On DEAP, the CogEmoNet MSE (3.654) is 4.793 and 1.085 less than the baseline (8.447) and hybrid CNN-LSTM (4.739) respectively. The CogEmoNet CCC (0.181) is 0.064 and 0.041 higher than the baseline (0.117) and hybrid CNN-LSTM (0.140) respectively. On DEFE+, the CogEmoNet MSE (4.220) is 5.682 and 1.582 less than the baseline (9.902) and hybrid CNN-LSTM (5.802) respectively. The CogEmoNet CCC (0.221) is 0.066

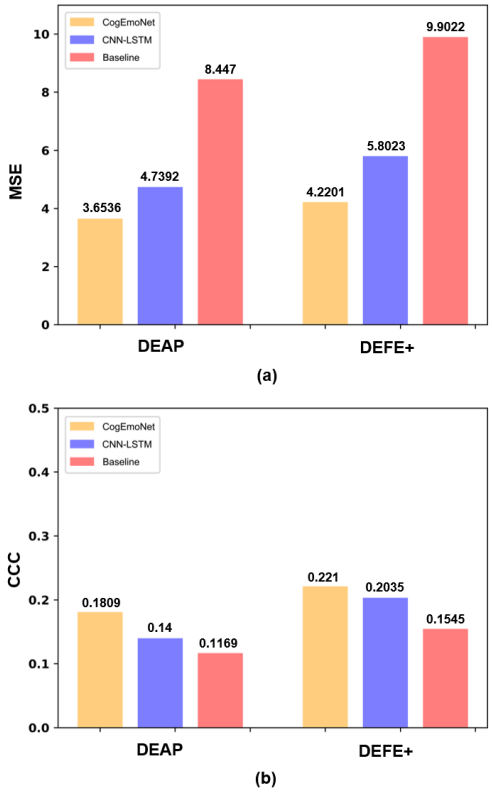


Fig. 10: MSE and CCC of dimensional emotion model

and 0.017 higher than the baseline (0.155) and hybrid CNN-LSTM (0.204) respectively. The effectiveness of our proposed CogEmoNet in dimensional emotion recognition was verified.

Table III and Table IV demonstrate the experiment results for all the performance metrics. The tables show that the CogEmoNet recognition framework is capable to achieve well detection results on different datasets. we also find that compared with the CK+, the recognition results obtained on DEFE+ and DEAP are lower. Obviously, the data in the CK+ contain a wide range of variations, for example, including participants from different nationalities. At the same time, posed or spontaneously induced may lead to the difference, most of the CK+ data are posed by the participants, while the DEFE+ and DEAP data are spontaneously induced. Moreover, the facial expression of the driver may be suppressed due to the driving tasks, which may be the reason for the difference between DEFE+ and DEAP recognition results. Furthermore, The performance of the CogEmoNet model on the CK+ dataset is not good enough in comparison with previous studies [23] [28]. This is mainly caused by different data processing methods and sampling methods. This study uses a fixed number of expression frames instead of peak expression frames for training and uses a person-independent sampling method for verification. These processes all increase the difficulty of model learning.

VI. Conclusion

In this paper, we proposed a cognitive-feature-augmented model to detect driver emotion based on facial expression and cognitive process characteristics. This model was proposed and implemented by combining emotion generation process theory and deep learning algorithms. CogEmoNet recognized driver emotions by simultaneously considering the drivers facial expression and cognitive process characteristics. To verify the performance of the CogEmoNet, This paper conducted driver's emotion data collection. The collected dataset included frontal facial videos from 40 drivers, their cognitive process characteristics, and subjective ratings on driver emotions. CBAM_ResNet34 and CNN-LSTM were also used to compare detection performance. The results show that the CogEmoNet detection architecture is capable of achieving well detection results for different databases on discrete emotion model and dimensional emotion model, respectively.

References

[1] A. Bhat, S. Aoki, and R. Rajkumar, "Tools and methodologies for autonomous driving systems," *Proc. IEEE*, vol. 106, no. 9, pp. 1700–1716, 2018.

[2] G. Li, Y. Yang, X. Qu, D. Cao, and K. Li, "A deep learning based image enhancement approach for autonomous driving at night," *Knowledge-Based Systems*, vol. 213, p. 106617, 2021.

[3] W. Li, B. Zhang, P. Wang, C. Sun, G. Zeng, Q. Tang, G. Guo, and D. Cao, "Visual-attribute-based emotion regulation of angry driving behaviours," *IEEE Intell. Transp. Syst. Mag.*, vol. DOI:10.1109/MITS.2021.3050890, 2021.

[4] 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. DOI:10.1109/TAFFC.2021.3063387, 2021.

[5] D. McDuff and M. Czerwinski, "Designing emotionally sentient agents," *Commun. ACM*, vol. 61, no. 12, pp. 74–83, 2018.

[6] M. Braun, R. Chadowitz, and F. Alt, "User experience of driver state visualizations: A look at demographics and personalities," in *IFIP Conf. Hum. Comput. Interact.*, pp. 158–176, Springer, 2019.

[7] S. Li, T. Zhang, N. Liu, W. Zhang, D. Tao, and Z. Wang, "Drivers attitudes, preference, and acceptance of in-vehicle anger intervention systems and their relationships to demographic and personality characteristics," *Int. J. Ind. Ergon.*, vol. 75, p. 102899, 2020.

[8] X. Wang, Y. Liu, F. Wang, J. Wang, L. Liu, and J. Wang, "Feature extraction and dynamic identification of drivers emotions," *Transp. Res. Pt. F-Traffic Psychol. Behav.*, vol. 62, pp. 175–191, 2019.

[9] G. Li, Y. Chen, D. Cao, X. Qu, B. Cheng, and K. Li, "Extraction of descriptive driving patterns from driving data using unsupervised algorithms," *Mech. Syst. Signal Process.*, vol. 156, p. 107589, 2021.

[10] P. Wan, C. Wu, Y. Lin, and X. Ma, "On-road experimental study on driving anger identification model based on physiological features by roc curve analysis," *IET Intell. Transp. Syst.*, vol. 11, no. 5, pp. 290–298, 2017.

[11] B. G. Lee, T. W. Chong, B. L. Lee, H. J. Park, Y. N. Kim, and B. Kim, "Wearable mobile-based emotional response-monitoring system for drivers," *IEEE T. Hum.-Mach. Syst.*, vol. 47, no. 5, pp. 636–649, 2017.

[12] L. Malta, C. Miyajima, N. Kitaoka, and K. Takeda, "Analysis of real-world driver's frustration," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 1, pp. 109–118, 2010.

[13] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proc. Int. Conf. Multimodal Interfaces*, pp. 205–211, 2004.

[14] J. A. Groeger, *Understanding driving: Applying cognitive psychology to a complex everyday task*. Psychology Press, 2000.

[15] T. Lajunen, D. Parker, and H. Summala, "The manchester driver behaviour questionnaire: a cross-cultural study," *Accid. Anal. Prev.*, vol. 36, no. 2, pp. 231–238, 2004.

[16] T. Brosch, K. R. Scherer, D. M. Grandjean, and D. Sander, "The impact of emotion on perception, attention, memory, and decision-making," *Swiss Med. Wkly.*, vol. 143, p. w13786, 2013.

[17] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, 2009.

[18] P. Ekman, "An argument for basic emotions," *Cogn. Emot.*, vol. 6, no. 3-4, pp. 169–200, 1992.

[19] J. A. Russell, "A circumplex model of affect.," *J. Pers. Soc. Psychol.*, vol. 39, no. 6, p. 1161, 1980.

[20] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Curr. Psychol.*, vol. 14, no. 4, pp. 261–292, 1996.

[21] J. J. Gross and R. W. Levenson, "Emotion elicitation using films," *Cogn. Emot.*, vol. 9, no. 1, pp. 87–108, 1995.

[22] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *J. Behav. Ther. Exp. Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[23] D. Liang, H. Liang, Z. Yu, and Y. Zhang, "Deep convolutional bilstm fusion network for facial expression recognition," *Vis. Comput.*, vol. 36, no. 3, pp. 499–508, 2020.

[24] K. Li, Y. Jin, M. W. Akram, R. Han, and J. Chen, "Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy," *Vis. Comput.*, vol. 36, no. 2, pp. 391–404, 2020.

[25] I. Gogić, M. Manhart, I. S. Pandžić, and J. Ahlberg, "Fast facial expression recognition using local binary features and shallow neural networks," *Vis. Comput.*, vol. 36, no. 1, pp. 97–112, 2020.

[26] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 4057–4069, 2020.

[27] Y. Wang, M. Li, C. Zhang, H. Chen, and Y. Lu, "Weighted-fusion feature of mb-lbpuh and hog for facial expression recognition," *Soft Comput.*, vol. 24, no. 8, pp. 5859–5875, 2020.

[28] D. Liu, X. Ouyang, S. Xu, P. Zhou, K. He, and S. Wen, "Saanet: Siamese action-units attention network for improving dynamic facial expression recognition," *Neurocomputing*, vol. 413, pp. 145–157, 2020.

[29] R. S. Lazarus and R. S. Lazarus, *Emotion and adaptation*. Oxford University Press on Demand, 1991.

[30] J. J. Gross, *Handbook of emotion regulation*. Guilford publications, 2013.

[31] C. D. Wickens, J. G. Hollands, S. Banbury, and R. Parasuraman, *Engineering psychology and human performance*. Psychology Press, 2015.

[32] R. Breuer and R. Kimmel, "A deep learning perspective on the origin of facial expressions," *arXiv preprint arXiv:1705.01842*, 2017.

[33] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 2983–2991, 2015.

[34] D. H. Kim, W. J. Baddar, J. Jang, and Y. M. Ro, "Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition," *IEEE Trans. Affect. Comput.*, vol. 10, no. 2, pp. 223–236, 2017.

[35] W.-S. Chu, F. De la Torre, and J. F. Cohn, "Learning spatial and temporal cues for multi-label facial action unit detection," in *2017 12th IEEE Int. Conf. Autom. Face Gesture Recognit.(FG 2017)*, pp. 25–32, IEEE, 2017.

[36] B. Hasani and M. H. Mahoor, "Facial expression recognition using enhanced deep 3d convolutional neural networks," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit. Workshops*, pp. 30–40, 2017.

[37] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vision (ECCV)*, pp. 3–19, 2018.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 770–778, 2016.

[39] Z. Zhang and M. R. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *32nd Conf.Neural Inf. Process. Syst. (NeurIPS)*, 2018.

[40] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation," *BMC genomics*, vol. 21, no. 1, pp. 1–13, 2020.

[41] C. R. Rao, "Some comments on the minimum mean square error as a criterion of estimation.," tech. rep., PITTSBURGH UNIV PA INST FOR STATISTICS AND APPLICATIONS, 1980.

[42] C. E. Izard, *Patterns of emotions: A new analysis of anxiety and depression*. Academic Press, 2013.

[43] S. Taamneh, P. Tsiamyrtzis, M. Dcosta, P. Buddharaju, A. Khatri, M. Manser, T. Ferris, R. Wunderlich, and I. Pavlidis, "A multimodal dataset for various forms of distracted driving," *Scientific data*, vol. 4, no. 1, pp. 1–21, 2017.

[44] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *Eur. Conf. Comput. Vision*, pp. 87–102, Springer, 2016.

[45] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit.-Workshops*, pp. 94–101, IEEE, 2010.

[46] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, 2011.

[47] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, 2016.

[48] D. K. Jain, P. Shamsolmoali, and P. Sehdev, "Extended deep neural network for facial emotion recognition," *Pattern Recognit. Lett.*, vol. 120, pp. 69–74, 2019.

[49] M. A. Takalkar, S. Thuseethan, S. Rajasegarar, Z. Chaczko, M. Xu, and J. Yearwood, "Lgattnet: Automatic micro-expression detection using dual-stream local and global attentions," *Knowl.-Based Syst.*, vol. 212, p. 106566, 2021.

[50] E. Paul, "Emotions revealed: recognizing faces and feelings to improve communication and emotional life," *NY: OWL Books*, 2007.

[51] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

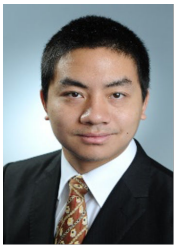
Wenbo Li received the B.S., M.Sc. and Ph.D. degree in automotive engineering from Chongqing University, Chongqing, China, in 2014, 2017 and 2021, respectively. He is also a visiting Ph.D. student at the Waterloo Cognitive Autonomous Driving (CogDrive) Lab at University of Waterloo, Canada. His research interests include smart cockpit, intelligent vehicle, human emotion, driver emotion detection, affective computing, emotion regulation, human-machine interaction, brain computer interface.

Guanzhong Zeng received the B.Sc. degree in Mechatronics Engineering and M.Sc. degree in Vehicle Engineering from Chongqing University in Chongqing, China, in 2018 and 2021, respectively. His research interests include facial expression recognition, fine-grained image recognition and gaze estimation.

Juncheng Zhang received the B.Sc. degree in Vehicle Engineering from Beijing Forestry University in Beijing, China, in 2016. He is currently pursuing the M.Sc degree in Chongqing University, Chongqing, China. His research interests include emotion recognition and natural language processing.

Yan Xu received his B.S. degree in mechanical engineering at University of Science and Technology Beijing, Beijing, China, in 2016. He is currently pursuing the Ph.D. degree in mechanical engineering at University of Science and Technology Beijing, Beijing, China. His research interest includes the machine learning, computer vision, intelligent vehicle.

Yang Xing received his Ph. D. degree from Cranfield University, UK, in 2018. He is currently a Lecture with Cranfield University. Before joining Cranfield, he was a Research Associate with the Department of Computer Science, University of Oxford, UK, from 2020 to 2021, and a Research Fellow with the School Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore, from 2019 to 2020. His research interests include machine learning, human behavior modeling, intelligent multi-agent collaboration, and intelligent/autonomous vehicles. He received the IV2018 Best Workshop/Special Issue Paper Award. Dr. Xing serves as a Guest Editor for IEEE IoT, IEEE ITSM, and Frontiers in Mechanical Engineering. He is also an active reviewer for IEEE ITS, TVT, TIE, and IEEE/ASME Transactions on Mechatronics, etc.



Rui Zhou received his B.Sc. degree in automobile engineering from Tongji University and the M.Sc degree in automobile engineering from Technical University of Braunschweig in 2010 and 2014. He is currently working as Research and Development Director at Waytous Inc. in China. He has served as software engineer and test engineer for Daimler AG. in Stuttgart and Ford-Werke GmbH. in Cologne. His research interests include autonomous vehicle, test area for intelligent-connected vehicle and functional safety.

Gang Guo received the Ph.D degrees in mechanical engineering from Chongqing University, Chongqing, China, in 1994. He is currently the professor at the Department of Mechanical and Vehicle Engineering, Chongqing University. He has authored and co-authored over 100 refereed journal and conference publications. His research interests include human-machine interaction, user experience, smart cockpit, intelligent vehicle, brain computer interface,and intelligent manufacturing.

Yu Shen received his master degree from University of Chinese Academy of Science in 2018 He is currently a PH.D student at School of Artifitical Intelligence, University of Chinese Academy of Science. His research interests includes Parallel Sensing, Computer vision and intelligent transportation systems.

Dongpu Cao (M08) received the Ph.D. degree from Concordia University, Canada, in 2008. He is the Canada Research Chair in Driver Cognition and Automated Driving, and currently an Associate Professor and Director of Waterloo Cognitive Autonomous Driving (CogDrive) Lab at University of Waterloo, Canada. His current research focuses on driver cognition, automated driving and cognitive autonomous driving. He has contributed more than 200 papers and 3 books. He was honored with the SAE Arch T. Colwell Merit Award in 2012, and three Best Paper Awards from the ASME and IEEE conferences. Dr. Cao serves as an Associate Editor for IEEE TVT,IEEE TIE, IEEE ITS, IEEE/ASME TRANSACTIONS ON MECHATRONICS, IEEE/(JAC)CAA and J DYN SYST-T ASME. He was a Guest Editor for VEHICLE SYSTEM DYNAMICS and IEEE SMCS. He serves on the SAE Vehicle Dynamics Standards Committee and acts as the Co-Chair of IEEE ITSS Technical Committee on Cooperative Driving.

Fei-Yue Wang (S87-M89-SM94-F03) received the Ph.D. degree in computer and systems engineering from Rensselaer Polytechnic Institute, Troy, NY, USA, in 1990. He joined The University of Arizona, Tucson, AZ, USA, where he became a Professor and the Director of the Robotics and Automation Laboratory and the Program in Advanced Research for Complex Systems. In 1999, he founded the Intelligent Control and Systems Engineering Center, Institute of Automation, Chinese Academy of Sciences (CAS), Beijing, China. In 2002, he participated in the development of the Key Laboratory of Complex Systems and Intelligence Science, CAS, as the Director, where he was also the Vice President for Research, Education, and Academic Exchanges at the Institute of Automation from 2006 to 2010. In 2011, he was named as the Director of the State Key Laboratory for Management and Control of Complex Systems, Beijing. His current research interests include methods and applications for intelligent and parallel systems, social computing, parallel intelligence, and knowledge automation. Dr. Wang was elected Fellow of the INCOSE, the IFAC, the ASME, and the AAAS. He was honored with the Best Paper Awards for his work from the IEEE ITS Society in 2012 and the IEEE CI Society in 2017, the Franklin V. Taylor Memorial Award in 2002, and the Andrew P. Sage Award from the IEEE SMC Society in 2019. He was also a recipient of the IEEE ITS Outstanding Application and Research Awards in 2009, 2011, and 2015, and the IEEE SMC Norbert Wiener Award in 2014.He was the President of the IEEE ITS Society from 2005 to 2007; the Chinese Association for Science and Technology, USA, in 2005; and the American Zhu Kezhen Education Foundation from 2007 to 2008. He was the Vice President of the ACM China Council from 2010 to 2011 and the Chair of the IFAC TC on Economic and Social Systems from 2008 to 2014 and from 2017 to 2023. He is the President of the IEEE Council on RFID and the Vice President of the IEEE SMC Society.He was the Founding EiC of International Journal of Intelligent Control and Systems from 1995 to 2000, IEEE ITSM from 2006 to 2007, and IEEE/(JAS)CAA from 2014 to 2017. He was the EiC of IEEE INTELLIGENT SYSTEMS from 2009 to 2012, IEEE TITS from 2009 to 2016, and IEEE TCSS from 2017 to 2020, and the Founding EiC of the Chinese Journal of Command and Control and the Chinese Journal of Intelligent Science and Technology. He received the Best Paper Awards for his work from the IEEE ITSS in 2012 and the IEEE Computational Intelligence Society in 2017, the Franklin V. Taylor Memorial Award in 2002, and the Andrew P. Sage Award from the IEEE Systems, Man, and Cybernetics Society (SMCS) in 2019. In 2007, he was a recipient of the National Prize in Natural Sciences of China and was awarded the Outstanding Scientist by the Association for Computing Machinery (ACM) for his research contributions in intelligent control and social computing. He was also a recipient of the IEEE ITS Outstanding Application and Research Awards in 2009, 2011, and 2015, and the IEEE SMC Norbert Wiener Award in 2014.