Journal of Electronic Imaging

JElectronicImaging.org

Human fatigue expression recognition through image-based dynamic multi-information and bimodal deep learning

Lei Zhao
Zengcai Wang
Xiaojin Wang
Yazhou Qi
Qing Liu
Guoxin Zhang

Lei Zhao, Zengcai Wang, Xiaojin Wang, Yazhou Qi, Qing Liu, Guoxin Zhang, "Human fatigue expression recognition through image-based dynamic multi-information and bimodal deep learning," *J. Electron. Imaging* **25**(5), 053024 (2016), doi: 10.1117/1.JEI.25.5.053024.

Journal of Electronic Imaging 25(5), 053024 (Sep/Oct 2016)

Human fatigue expression recognition through image-based dynamic multi-information and bimodal deep learning

Lei Zhao,[a] Zengcai Wang,[a,b,*] Xiaojin Wang,[a] Yazhou Qi,[a] Qing Liu,[a] and Guoxin Zhang[a]
[a]Shandong University, School of Mechanical Engineering, Vehicle Engineering Research Institute, No. 17923 of Jingshi Road, Jinan 250061, China
[b]Shandong University, School of Mechanical Engineering, Ministry of Education, Key Laboratory of High Efficiency and Clean Mechanical Manufacture, No. 17923 of Jingshi Road, Jinan 250061, China

Abstract. Human fatigue is an important cause of traffic accidents. To improve the safety of transportation, we propose, in this paper, a framework for fatigue expression recognition using image-based facial dynamic multi-information and a bimodal deep neural network. First, the landmark of face region and the texture of eye region, which complement each other in fatigue expression recognition, are extracted from facial image sequences captured by a single camera. Then, two stacked autoencoder neural networks are trained for landmark and texture, respectively. Finally, the two trained neural networks are combined by learning a joint layer on top of them to construct a bimodal deep neural network. The model can be used to extract a unified representation that fuses landmark and texture modalities together and classify fatigue expressions accurately. The proposed system is tested on a human fatigue dataset obtained from an actual driving environment. The experimental results demonstrate that the proposed method performs stably and robustly, and that the average accuracy achieves 96.2%. *© 2016 SPIE and IS&T* [DOI: 10.1117/1.JEI.25.5.053024]

Keywords: fatigue expression recognition; texture; landmark; bimodal learning; dynamic multi-information.

Paper 16416 received May 14, 2016; accepted for publication Sep. 15, 2016; published online Oct. 6, 2016.

1 Introduction

The National Highway Traffic Safety Administration[1] investigated 4010 drivers in the United States and found that 37% of the respondents admitted to experiencing fatigue while driving. Among the respondents who felt fatigue driving, 29% experienced fatigue driving in the past year and 10% experienced it in the past month. In Europe,[2] 10% to 20% of truck traffic accidents were caused by fatigue driving. Compared with those of overspeeding, drunk driving, and other dominant risk behaviors during driving, the identification and prevention of fatigue driving are difficult, and thus fatigue driving poses greater harm to human security. Therefore, examining effective methods for detecting human fatigue is essential to improving transport safety. Over the past 20 years, driving fatigue detection has become an increasing concern of the industry and the scientific community, and many achievements and industrial products have been developed. At present, the driving fatigue recognition system is an indispensable part of the advanced driver assistance system.

Many studies[3] have shown that physiological indexes change when a person feels drowsy. Thus, whether or not a driver is in a state of fatigue can be determined by detecting physiological signals. At present, mature physiological signal detection technologies are available for recognizing driver fatigue. The main physiological signals include electroencephalogram, electrocardiogram, and electrooculogram, among others. However, most physiological sensors are invasive and need to be pasted on the surface of the human body, thus causing the discomfort to the driver.

Marked changes in behaviors of operation are observed when a driver becomes fatigued. Therefore, fatigue can also be detected by the operational behaviors of drivers.[4] Behaviors of drivers such as steering wheel movements,[5] lateral position, driver's grip force on the steering wheel,[6] speed, turning angle, acceleration, changing course, gear changing, and braking are used to evaluate driver fatigue. Compared with physiological methods, those based on driver performance are nonintrusive and extracted easily and accurately. However, driver experiences, vehicle type, and driving conditions may affect the accuracy of identification.

People experiencing drowsiness exhibits certain facial or head behaviors that can be observed easily. Facial behaviors, which include eyelid movement, degree of eye openness, yawning, pupil movement, facial expression, and head rotation, can reflect the level of a person's drowsiness. Compared with the methods based on physiological signals and driver behaviors, the major advantage of the visual measures using image-based technologies is that facial behaviors can be extracted nonintrusively and are not subject to driving experience, vehicle type, and driving conditions. Furthermore, most image-based methods require only several cameras to capture the facial images of drivers, which occupy a small volume and entail low system cost to recognize driver fatigue.

Many image-based methods that use video cameras or optical sensors to capture facial geometry or textural features and fatigue cues of head posture have been proposed to recognize driver fatigue in the last decade. Furthermore, driver fatigue detection system products[7] have been developed and

*Address all correspondence to: Zengcai Wang, E-mail: wangzc@sdu.edu.cn

put into practice in recent years. Although an image-based method for fatigue recognition may include different aspects, recent studies on this topic can be divided into four main streams that consider eye state parameters, mouth state parameters, head posture parameters, and facial expression parameters.

1.1 *Methods Based on Eye Parameters*

The time and frequency of eye closure increase when people are drowsy. Therefore, researchers use eye state as the feature by which to judge fatigue. Dinges and Grace[8] proposed the percentage of eyelid closure (PERCLOS), which is the most reliable parameter to measure a person's fatigue level based on the dataset of the Federal Highway Administration. Furthermore, the best correlation was found between P80 of PERCLOS and the level of human fatigue. Most driving fatigue system products,[7] including the drowsy driver detection system designed by Johns Hopkins University, copilot system developed by Carnegie Mellon University, and A-PCS designed by Toyota, have been used in this manner since then. In addition to this approach, researchers have explored other driving fatigue detection techniques that use eye features. Liu and Wechsler[9] combined mean shift and Kalman filtering to track eye regions, and the motion features of the eyes were extracted to detect fatigue. Rong-Ben[10] extracted the textural features of drivers' eye regions through two-dimensional Gabor wavelet transform and used neural networks to classify the behavior of drivers' fatigue. Dong and Wu[11] detected the distance between the upper and lower eyelids to judge whether or not a driver is fatigued. Fan et al.[12] extracted local binary pattern (LBP) histograms of eyes and selected the most discriminative features to recognize human fatigue using the AdaBoost algorithm. Kim et al.[13] proposed a fuzzy logic system using I and K color information from the hue, saturation, and intensity and the cyan, magenta, yellow, and key color spaces of eye region to classify eye states. Ibrahim et al.[14] proposed an interdependent and adaptive scale mean shift algorithm that uses moment features to track and estimate the iris area to classify the state of the eye, and the test results demonstrated that the system improves nonrigid eye tracking performance. Song et al.[15] proposed a new feature descriptor called multiscale histograms of principal oriented gradients to detect whether the eyes are closed in a static face image. Cyganek and Gruszczyński[16] proposed a hybrid visual system for monitoring drivers' state of fatigue based on drivers' eye recognition using two cameras. In this system, a cascade of two classifiers is used to detect the eyes. The first classifier is used for detecting eye regions, and the second classifier is used for eye verification.

1.2 *Methods Based on Mouth Parameters*

People may yawn when they are fatigued. Therefore, the variation in shape or texture of the mouth can be extracted to judge whether or not a driver yawns. Rongben et al.[17] extracted the shape of the mouth to classify three different mouth states (i.e., normal, yawning, and talking states) using a back propagation neural network to recognize human drowsiness. Wang and Shi[18] applied the mouth height-to-width ratio to represent the openness of the mouth. Yawning can be recognized if the value of the ratio exceeds 0.5 in more than 20 frames. Yufeng and Zengcai[19] located the nose tip using directional integral projection and judged yawns by calculating the vertical distance between the nose tip and the midpoint of the chin.

1.3 *Methods Considering Head Posture*

The ability of the driver to control head movements is usually related to the level of fatigue. In certain studies in the past several years, head movement parameters were extracted to analyze the correlation with distraction and usually integrated with eye state parameters to recognize driver fatigue. Lee et al.[20] proposed two methods to estimate the head rotation of the driver to detect driver fatigue and distraction: an ellipsoidal head model instead of the cylindrical model to exactly determine a driver's yaw, and a normalized standard deviation and mean of the horizontal edge projection histogram to reliably estimate a driver's pitch. Oyini Mbouna et al.[21] continuously monitored the alertness states of a vehicle driver using eye state and head pose. The eye state contains the position of the pupil and PERCLOS, and the head pose includes pitch, yaw, and tilting. In Ref. 7, several vehicle-based noninvasive drivers' monitoring systems that consider the parameters of head posture, such as SmartEye, DDS, and FaceLAB system, are reviewed and evaluated.

1.4 *Methods Based on Facial Expression*

Facial expression recognition, a significant and challenging problem in the fields of pattern recognition and computer vision, attracted considerable attention in the last several years.[22–27] An increasing number of driver fatigue detection methods based on facial expression were also proposed. Gu and Ji[28] proposed human fatigue detection systems based on a dynamic Bayesian network and the general facial expression language (FACS) to recognize fatigue expression. Fan et al.[29] extracted the dynamic facial expression features using the Gabor multiorientation fusion histogram in five frames and used the AdaBoost algorithm to select the most discriminative features to recognize human fatigue. Five rules of the proposed feature extraction methods were adopted, and the experiments demonstrated that the average recognition rate could reach 99.3%. In Ref. 30, the pyramid histogram of oriented gradients and contourlet transform of facial region were extracted as discriminate features, and a random subspace ensemble (RSE) of linear perception classifier was then applied to classify three predefined fatigue expression categories. Zhang and Hua[31] extracted the LBP features of a face image and used the AdaBoost algorithm to select the most discriminative features to recognize human fatigue expression. Zhao et al.[32] combined Curvelet transform and Gabor wavelet transform to describe the fatigue expressions of drivers and used the RSE of support vector machines (SVMs) with polynomial kernel to classify three fatigue expression categories, namely, awake, moderate fatigue, and severe fatigue.

1.5 *Discussion and Contributions*

The parameters of eyes and mouth are widely adopted to recognize driver fatigue because they can reflect the facial behavior when a driver is drowsy. However, a single clue from the eyes or mouth overlooks numerous crucial features

that complement each other. Therefore several automotive driver monitoring systems based on a single clue usually perform poorly in real applications. The parameters of head posture are commonly used for integration with other features to recognize driver fatigue. However, the parameters have several limitations, such as driving habits, pavement condition, and individual differences. Therefore, in several drivers' monitoring systems that consider head posture, the parameters of head movement are usually adopted to evaluate the distraction rather than the fatigue of drivers. Systems based on facial expression can capture multiple cues of facial behaviors that are not related to the driving environment as well as the driver's habits to improve the accuracy and robustness of fatigue recognition. However, most driver fatigue detection methods considering facial expression focus on the means of analyzing static facial images (e.g., static-based method). Static-based methods require only one shot to analyze facial expressions and ignore several important dynamic expression features. Furthermore, most features used in previously proposed methods are texture (e.g., Gabor and LBP) of the face region. However, most textural features are susceptible to appearance changes (e.g., illumination and colors of skin), and these handcrafted features are cumbersome, time-consuming, and require complex computing.

To solve the above-mentioned problem, we extracted landmarks of eyes and mouth and the texture of eye regions that complement each other as dynamic multi-information from the image sequence to detect fatigue expression. The landmark is usually represented as xy-coordinates of facial points, and texture is represented using real-valued and dense pixel intensities. The relationship across two modalities is difficult to discover. The two modalities carry different types of information at a low-level but are correlated at a high-level for fatigue expression recognition. Therefore, a bimodal deep learning algorithm based on a stacked autoencoder (SAE) is established to learn the fusion information of two modalities. First, two SAEs are trained for landmark and texture, respectively. Then, the two pretrained networks are combined by learning a joint layer on top of the networks to obtain the cross-modal representation. Finally, a softmax regression is added on the top layer of the network to perform classification. Using the algorithm mentioned earlier, we can determine the joint latent and abstract representations that capture the relationship between landmark and texture modalities to boost the performance of fatigue expression recognition. The main contributions of this paper are as follows:

a. A driver fatigue recognition system based on dynamic multi-information of facial image sequences and deep learning algorithm.
b. A framework of human fatigue expression recognition that integrates landmark with the texture of eye regions in image sequences.
c. A bimodal deep neural network (DNN) architecture based on an SAE.

The remainder of this paper is structured as follows. Section 2 describes the approaches we proposed in detail. Section 3 presents the experiments used to evaluate the performance of the proposed method. Finally, Sec. 4 concludes the paper and proposes further works.

2 Methods

The proposed approach includes face detection and facial point location (preprocessing), feature (modality) extraction, and bimodal deep learning. An exhaustive illustration of this method is shown in Fig. 1.

2.1 *Preprocessing*

When an image sequence is given to a fatigue expression recognition system, the facial regions are detected and cropped as a preprocessing step. The real-time Viola–Jones face detector,[33] which is the most commonly used approach in many areas, such as face detection, recognition, and expression analysis, is adapted to crop the face region in this paper. The Viola–Jones face detector consists of a cascade of classifiers that employs the Haar feature trained by AdaBoost. The Haar feature is based on integral image filters that can be computed simply and rapidly at any location and scale.

In this paper, the constrained local model (CLM) based on incremental formulation[34] is employed to detect and track facial points. The original CLM framework is a patch-based method, and the face can be represented by a series of image

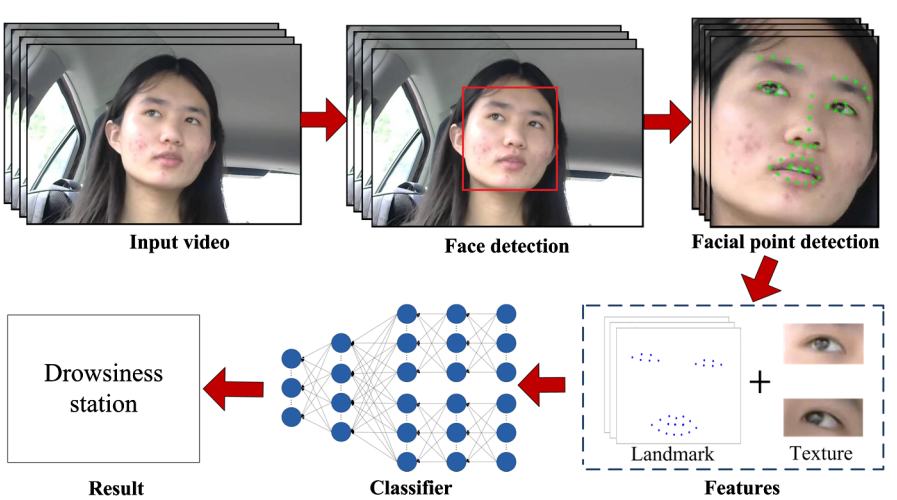

Fig. 1 Proposed framework for fatigue expression recognition.

patches cropped from the location of fiducial points. The patch-expert of each point is trained using a linear SVM and positive and negative patches, and it is used as a detector to update the location of the point. Instead of dealing with updating the shape model, the CLM based on incremental formulation focuses on incrementally updating the function that can map facial texture to facial shape, and automatically construct a robust and discriminative deformable model that outperforms state-of-the-art CLM alignment frameworks.[34] Figure 2 shows the location results using this method. As shown in Fig. 2, the facial fiducial points can be correctly positioned in a variety of possible situations, such as facial accessories [Figs. 2(a) and 2(e)], occlusions [Fig. 2(b)], head rotation [Figs. 2(b) and 2(f)], and illumination [Fig. 2(d)].

2.2 *Dynamic Feature (Modality) Extraction*

According to previous studies, the state of the mouth and eyes can indicate fatigue expression. Therefore, the xy-coordinates of points on the contours of the eyes and mouth are extracted for fatigue expression recognition from image sequence. However, certain interferences (e.g., illuminations, face expression, and head rotation) may cause an inaccurate location of the points on the contours of the eyes [Figs. 2(b) and 2(c)]. Moreover, compared with the changes in the xy-coordinate of the points on the mouth, the variations in the xy-coordinate of the points on the eyes are considerably smaller, and thus more susceptible to interferences. Accordingly, landmark modality is difficult to use in classifying the fatigue expression more accurately in the actual driving environment. Thus, the image regions of the eyes are cropped according to the position of the points on the eye corner and integrated with the landmark to boost the performance of fatigue expression recognition.

In a short time (several frames), fatigue and alert state tend to produce apparently identical expressions, such as eye closure versus blinking and yawning versus speaking. In practice, when a driver is awake, the facial movement speed of talking or blinking tend to be faster and change more frequently than the facial behavior during fatigue. If the features are extracted in a relatively short period of time for recognition, the features between fatigued and alert may to be confused. Therefore, the driver's facial features should be extracted for a longer period of time to obtain sufficient information and reflect the differences between the alert state and the fatigue state. In this paper, the length of image sequences extracted for recognition fatigue is 1 s. As previously mentioned, facial movements are usually slow and last for a long time when a driver is fatigued. However, high temporal resolution can result in an increased amount of calculation and data dimension, which is disadvantageous to recognition accuracy. Consequently, the frame rate of the image sequence used to extract the two modalities is 12 frames/s in our study.

2.3 *Bimodal Deep Neural Network Based on Landmark and Texture*

2.3.1 *Stacked autoencoder neural network*

The concept of deep learning is based on artificial neural networks. A DNN can form an abstract high-level representation of attribute categories or features by combining low-level features to discover the distributed representation of data. An unsupervised greedy algorithm used to solve the problem of deep structure-related optimization was first proposed by Hinton and Salakhutdinov,[35] and their approach opened up a research upsurge in deep learning. At present, deep learning has been applied to different fields,[36] including image recognition,[37] speech recognition,[38] potential drug molecules,[39] and analysis of particle accelerator data[40]. In addition, different kinds of DNN structures are widely used, such as SAE,[41] deep belief network (DBN),[35] convolutional neural networks,[37] and recurrent neural networks (RNN)[42]. In this paper, SAE is used to construct a bimodal DNN.

An SAE is a neural network that contains the multiple layers of autoencoders (AE).[43] The hidden layer of each AE is wired to the input layer of the next AE. As shown in Fig. 3, AE consists of the encoder and decoder. The encoder maps the input x to a hidden representation y, and the decoder decodes the hidden representation back to reconstruct the version of x. Therefore, AE is an unsupervised algorithm that uses the back propagation (BP) algorithm and makes the output values $\hat{x}$ equal to the input values x. The activation function between the two layers is as follows:

$$y = f(z) = \sigma(Wx + b), \tag{1}$$

where $y \in R^m$ is the hidden layer, $x \in R^n$ is the input of AE, $W \in R^{m \times n}$, and b is the bias of the network. σ represents a nonlinear function, such as tanh, sigmoid, or rectified linear unit function.

To obtain the parameters of the encoder, the loss function of AE, which represents the reconstruction error, is minimized using the stochastic gradient descent algorithm

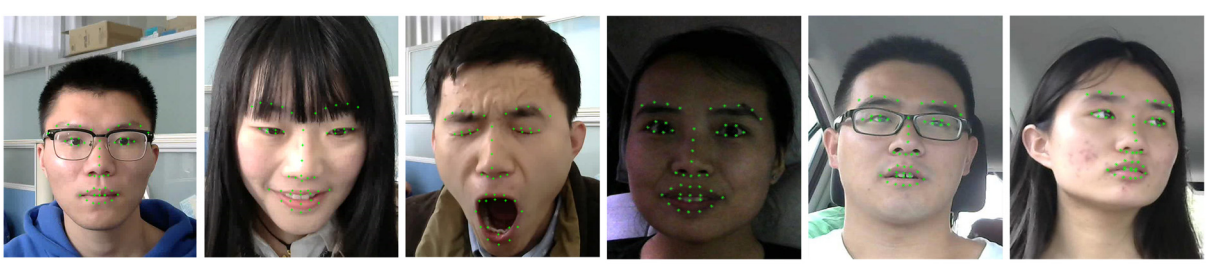

Fig. 2 Typical results of the proposed facial point detector for samples in different situations. (a) Facial accessory; (b) facial occlusion; (c) facial behavior; (d) illumination variation; (e) reflection on eyeglasses; and (f) head rotation.

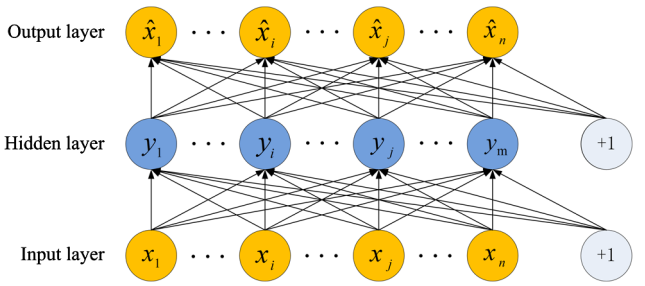

Fig. 3 AE structure.

$$\min_{W,b,k} L(x,W,b,h) = \frac{1}{2n}\sum_{k=1}^{n} \|\hat{x}_k - x_k\|_2^2, \tag{2}$$

$$y = \sigma(Wx + b), \tag{3}$$

$$\hat{x} = \sigma(Wy + h), \tag{4}$$

where h is the decoder bias and $L(x,W,b,h)$ is the loss function that represents the reconstruction error. $\hat{x} \in R^n$ is the output of AE.

Each layer of DNN is pretrained using an AE algorithm to consequently obtain the initial parameters.[41] Finally, the BP algorithm is used to fine-tune the whole network to obtain the final parameters of DNN for classification.[41]

2.3.2 *Bimodal deep neural network*

The main idea of a multimodal model is that two separate models are used to learn low-level representations of each modality to obtain the higher-level representations that are correlated with each other. Ngiam et al.[44] proposed a multimodal algorithm to learn the joint representation from the image and audio modalities input using DBN. Both modalities were used to learn the representation, but only one modality was applied for fine-tuning to find the correlation between the image and audio modalities. The results show that AE can reconstruct both image and audio modalities when only one modality is presented for supervised learning. Inspired by Ref. 44, Srivastava and Salakhutdinov adopted multimodal DBN[45] and multimodal DBM[46] to learn the joint representation of text and image modalities. Different forms multimodal DNNs[47,48] were proposed for learning the correlation between texture and landmark modalities for facial expression reorganization. In this paper, an SAE-based bimodal deep learning algorithm is proposed to recognize human fatigue expression by learning the correlation of the texture of the eye region and facial landmarks. The structure of the proposed network is shown in Fig. 4.

If we directly input the xy-coordinates of the facial points and pixel intensities of the image into AE in Eq. (1) after feature extraction, all nodes of the visible layer are connected with the nodes of the hidden layer. As mentioned earlier, specific properties of different modalities are ignored, and the correlations among different modalities cannot be well learned and represented. To overcome this limitation, we proposed a bimodal DNN structure as shown in Fig. 4. The bimodal learning algorithm contains four components, namely, two unimodal SAEs, a joint layer, and an output layer (i.e., the classifier). First, each unimodal SAE is pretrained separately by completely unsupervised learning, which requires a large number of training data. By accounting for different kinds of input data, the type of lower layer could be different, whereas the representations of the final hidden layer at the end of each unimodal SAE are of the same type. Then, a joint hidden layer is added on top of the two unimodal SAEs to learn the fusion representation of landmark and texture modalities using the AE algorithm (Fig. 4). Afterward, a joint latent and abstract representation that captures the relationship between landmark and texture modalities in the semantic layer is learned to boost the performance of fatigue expression recognition. Finally, the BP algorithm is used to fine-tune the network to obtain the optimal weights after adding a classifier on top of the joint

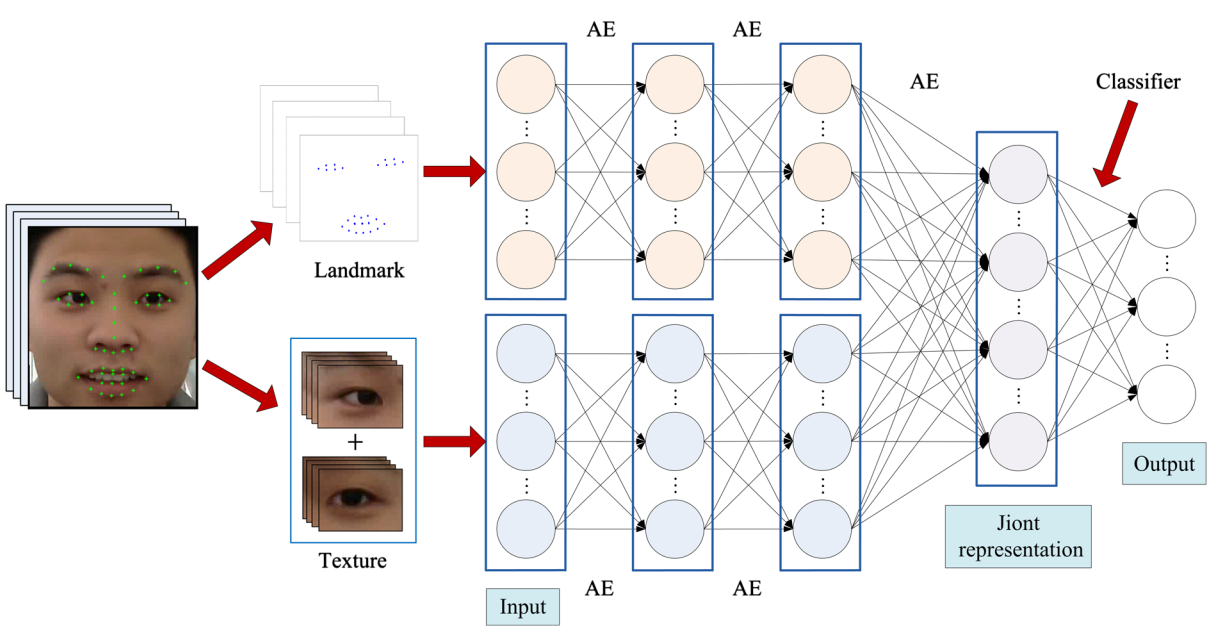

Fig. 4 Overall structure of the bimodal DNN.

layer to build the whole bimodal DNN. The following formulas describe the gradient descent method of the proposed network in the process of fine-tuning:

$$\begin{cases} \Delta W_{i,j}^{l} = \Delta V_{i,j}^{l} \cdot S_{i,j}^{l} & 1 \le l < d \\ \Delta W_{i,j}^{l} = \Delta V_{i,j}^{l} & d \le l < n \end{cases}, \tag{5}$$

where

$$\Delta V^{l} = \delta^{l+1} \cdot (a^{l})^{T}, \tag{6}$$

$$\delta_{i}^{l} = \left(\sum_{i=1}^{n_1} W_{i,j}^{l} \cdot \delta_{i}^{l+1} \right) \cdot f'(z^{l}), \tag{7}$$

$$\delta_{i}^{N} = -(Y_{i} - a_{i}^{N}) \cdot f'(z_{i}^{N}), \tag{8}$$

where $W_{i,j}^{l}$ is the corresponding weight of layer l and $\Delta W_{i,j}^{l}$ is the partial derivative of the loss function L on $W_{i,j}^{l}$. S^{l} is a modality binary matrix of the l'th layer, which is the same as the dimension of the weight matrix. δ_{i}^{l} is the error term of the node i in layer l, and δ_{i}^{N} is the error term of the node i in output layer. Y_{i} is the i'th value of the training label vector. a^{l} is the activation of layer l, and a_{i}^{N} is the i'th value of the output layer. z^{l} is a linear mapping vector of the activation value in layer $l-1$, i.e., $z^{l} = W^{l-1}a^{l-1} + b^{l-1}$. N is the number of layers in the bimodal DNN, and d is the number of layers in the unimodal DNN.

As shown in Eqs. (5)–(8) and Fig. 4, the gradient descent algorithm of the proposed network is not same as that of classical neural networks. This difference is due to the first d layers containing two unimodal SAEs, and the units of the two adjacent layers are not mapped to each other. Therefore, the matrix of the partial derivative ΔW^{l} $(l < d)$ needs to be multiplied by a binary matrix S^{l}, which can make the corresponding values in ΔW^{l} equal to 0.

3 Experiments

To evaluate the effectiveness of the proposed method, we established two datasets used in conducting the experiments. First, the detailed information on the two datasets is introduced. Second, we present the procedure of data augmentation and preprocessing. Third, the bimodal network architecture and training method are described in detail. Finally, the experimental results are provided to prove the effectiveness of the proposed method.

3.1 *Datasets*

Many public facial datasets, including facial expression or emotion recognition face recognition, and face detection, can be obtained easily. However, an image-based public fatigue recognition dataset to test the proposed method does not exist. Moreover, the datasets used in certain studies are extremely small and involves several subjects only. Therefore, two datasets are built using a high-definition camera to train and test the validity of the proposed method. One dataset that contains a large number of samples is used for the unsupervised pretraining of the DNN and the other obtained from an actual driving environment is used for training (fine-tuning) and testing. In recent studies, different classification criteria have been proposed to classify the degree of fatigue.[29–32] The criteria proposed in Refs. 30 and 32, including alert, moderate fatigue, and severe fatigue, adopted in our datasets.

3.1.1 *Training (fine-tuning) and testing dataset*

Considering the danger of the experiment while driving, we installed a high-definition camera in front of the passenger seat on the right of the driver to capture the passenger's image sequences, which are similar to the driver image sequences in facial behavior, head movement, and scenario, among others. The videos of 16 subjects are captured to obtain the facial fatigue expressions of each subject in a mutative scenario when the car is in motion. The age of the subjects ranges from 20 to 55 years, ~25% of the subjects are female, and half of the subjects wear glasses. Facial behaviors of each subject, including head rotation, different facial expression, and talking in an alert state, are spontaneous. All subjects were subjected to 6 to 8 h of straight sleep deprivation to induce fatigue before video capture. The ground truth of each subject's state is annotated by a researcher who continuously monitors each subject's facial behavior changes during video capture. Then, all videos are converted to AVI format compressed by MPEG-4. Image sequences of 1-s length containing different states of subjects are extracted from videos according to the human annotation and different facial expression (e.g., alert: talking, staring straight ahead, or looking around; moderate fatigue: yawn; severe fatigue: closing eyes gradually). Figure 5 shows the sample frames of the dataset. Finally, 419 image sequences are obtained as shown in Table 1. As shown in the sample images, various illumination conditions and head rotations are observed in the datasets.

3.1.2 *Pretraining dataset*

To obtain additional subjects easily to pretrain the DNN and enforce the ability of classifying unseen data accurately, a web camera is installed on the computer monitor and captures the facial image sequences of the person sitting in front of the camera. The videos of 30 subjects are captured

Fig. 5 Sample images in the fine-tuning and testing dataset.

Table 1 Number of training (fine-tuning) and testing datasets.

Classification	Alert	Moderate fatigue	Severe fatigue	Total
Number	225	88	106	419

Table 2 Number of pretraining datasets.

Classification	Alert	Moderate fatigue	Severe fatigue	Total
Number	445	192	205	842

indoors to obtain facial fatigue expressions of each subject. The age of the subjects ranges from 23 to 45 years, ~30% of the subjects are female, and seven subjects wear glasses. Except for the scenario, the methods of face image acquisition and processing in this dataset are the same as those in the fine-tuning and testing dataset, which is proposed in Sec. 3.1.1. Finally, 842 image sequences are obtained as shown in Table 2. Figure 6 shows sample frames of alert, moderate fatigue, and severe fatigue from six subjects. The sample images show that the illuminations conditions indoors are more stable than those in the automobile driving environment.

3.2 *Dataset Augmentation*

To classify unseen data accurately, a number of training data that cover various situations are required. However, the pretraining dataset and the fine-tuning dataset provide a limited number of image sequences. This fact can make a DNN easily overfit as a typical DNN contains many parameters. This problem requires various dataset augmentation techniques to solve. In this paper, a data augmentation method proposed in Ref. 48 is adapted to augment training data. First, each image in the training videos is horizontally flipped. The original images and flipped images are rotated by each angle in $\theta = (-15 \text{ deg}, -10 \text{ deg}, -5 \text{ deg}, 5 \text{ deg}, 10 \text{ deg}, 15 \text{ deg})$. This method can render the network robust against the slight in-plane head rotation of the subjects. Identical to the augmentation of image sequences, the facial landmark points are horizontally flipped. Rotating a landmark is constructed as follows:

$$[\tilde{x}_i^k, \tilde{y}_i^k]^T = R_i^k [x_i^k, y_i^k]^T, \tag{9}$$

where $\tilde{x}_i^k$ and $\tilde{y}_i^k$ are the i'th rotated xy-coordinates using R_i^k. R_i^k is the rotation matrix of angle $\theta(k)$ for landmark coordinates. Finally, we obtain 14 times more data

- Original data (1).
- Horizontally flipped data (1).
- Rotated data with six angles (6).
- Flipped versions of the rotated data (6).

3.3 *Data Preprocessing*

As previously mentioned, the augmented dataset contains both image (texture) and landmark modalities. These two modalities represent different properties of the facial behavior. As introduced in Ref. 49, data preprocessing is a critical steps before training network. Accordingly, two modalities are preprocessed as follows before being fed into the bimodal DNN.

The original image sequences are processed into the images with 8-bit precision for grayscale values for our study. The patches of eyes cannot cover the bridge of the nose or brows. The width of a rectangle of each eye patch is twice times as large as the distance between the two corners of the eye, and the height of a rectangle of each eye patch is the same as the distance between the two corners of the eye. The center point of the patch is located at the midpoint of the two corners. Then, these patches are further down sampled to 10×20 to reduce the data dimension and the complexity of computation for learning and classification. Finally, the patches extracted from one image sequence are concatenated into a row vector. The value of elements in the vector is the pixel value of grayscale image patches and the size of the vector is $200 \times 2 \times 12 = 4800$. Finally, the values of all vectors are scaled to $[-1,1]$.

To improve recognition accuracy and reduce the dimension of the input vector, points of the eyes and mouth are selected from all points on the face. The coordinates of these points cannot be appropriately used directly as an input to the deep network because the data are not normalized. Therefore, we subtract the xy-coordinate of the nose tip from the xy-coordinates of the point on the eyes and mouth to obtain the relative xy-coordinates of the eyes and mouth, which can be calculated as follows:

$$\Delta X_i^t = X_i^t - X_n^t, \tag{10}$$

$$\Delta Y_i^t = Y_i^t - Y_n^t, \tag{11}$$

where X_i^t and Y_i^t are the x and y coordinates of the i'th point on the eyes and mouth in time t, and X_n^t and Y_n^t are the x and y coordinates of the of the nose tip in time t, respectively. Then, similar to texture, the relative xy-coordinates of the facial points from one image sequence is concatenated

Fig. 6 Sample images in the pretraining dataset.

into a row vector, and the size of the resultant vector is $30 \times 2 \times 12 = 720$ for the dataset we established. Finally, the vector values of the landmark are also scaled to $[-1,1]$.

3.4 *Bimodal Network Architecture*

The landmark-based SAE consists of 720 visible units and 100 hidden units, followed by another layer of 100 hidden units. The texture-based SAE consists of 4800 visible units and 500 hidden units, followed by another layer of 500 hidden units. The joint hidden layer contains 600 units, and a softmax regression is added on the top of the network for classification.

3.5 *Validation Strategy*

In this paper, subject-based eightfold cross-validation strategy is applied in the experiments we conduct. In each validation, two subjects of the whole samples are selected to form the testing set, and the rest forms the training samples. The network was trained and tested for eight times, and each subject cannot be selected as a testing set repeatedly. For training and testing dataset, the augmented fine-tuning data are about $360 \times 14 = 5040$, which are 14 times of the original fine-tuning data and include 14 subjects. For pretraining dataset, the number of pretraining data is $842 \times 14 = 11{,}788$.

3.6 *Experiment Results*

3.6.1 *Unimodality versus bimodality*

The main idea of our method is to integrate facial landmarks with the eye texture for fatigue expression recognition. Accordingly, the performance of models with unimodality and bimodality should be respectively compared. Table 3 presents the experimental results of the three models. The accuracies of fatigue expression recognition demonstrate that the method using bimodal learning algorithm is more reliable than that using landmark or texture. We can observe that the method using texture modality shows the poorest performance of the three methods (84.0%) because the texture modality only represents the changes in the eye texture, whereas the changes in the mouth also contain important information for fatigue expression classification. Although the landmark modality achieves higher recognition accuracy (94.5%), certain interferences or positioning errors of the position of the eye contour points decrease the recognition rate. Consequently, the combination of two modalities yields the best recognition rate (96.2%) as shown in Table 3.

Table 3 Comparison of the algorithms with unimodality and bimodality.

Inputs	Alert (%)	Moderate fatigue (%)	Severe fatigue (%)	Average (%)
Texture	95.6	50.0	87.7	84.0
Landmark	98.7	90.9	88.7	94.5
Bimodality	99.1	90.9	94.3	96.2

3.6.2 *Comparison of the algorithms with and without autoencoder*

In this section, algorithms with and without pretraining are compared. The comparison results of the two methods are shown in Fig. 7. The results show that when the unsupervised pretraining is added to the learning process of the network, the accuracy of fatigue recognition improves. BP algorithm is used to minimize the output error using gradient descent algorithms. The initial values of the weights and bias of the network are usually random, and they may cause poor local optima if the network contains an excessive number of layers. If pretraining is adapted to initialize the weights and bias of each layer, the parameters of a deep network will not present the risk of falling into the local optima and can be fine-tuned to reach the optima using the BP algorithms after pretraining.

3.6.3 *Comparison of the algorithms with different input methods*

In this experiment set, we perform a comparison of the algorithms with different input methods. As shown in Table 4, fatigue expression recognition is conducted by first converting landmark and texture into a vector as input. Therefore, the network consists of 5520 visible units and all units between two adjacent layers are connected with each other. The recognition results are shown in the second row of Table 4. Afterward, as described in Sec. 3.4, landmark and texture are treated separately. Both parameters are regarded as two different modalities and fed into the bimodal DNN that we established. The recognition results are illustrated in the third row of Table 4. The algorithm pretraining the two modalities separately slightly outperforms the first algorithm. However, the advantage of the proposed method is not evident probably because the training data are relatively small, and make the network fall into overfit to degrade the performance of the proposed algorithm.

3.6.4 *Different parameters of the joint representative layers*

In this experiment set, we perform a comparison of the algorithms with different numbers of the joint representative layers and the number of units. The recognition results are shown in Fig. 8. As shown in the figure, the model is not sensitive to the selection of two parameters of the joint

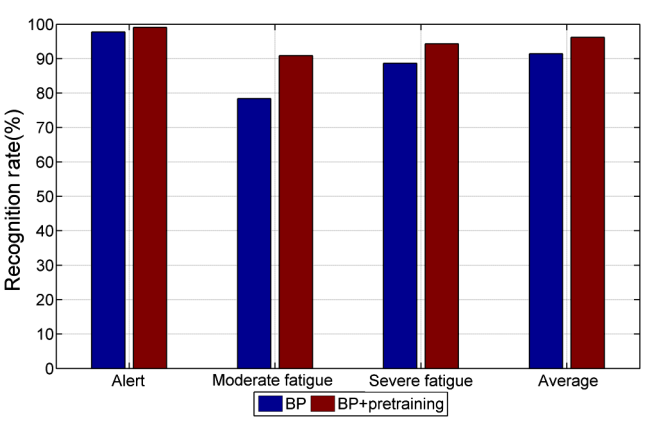

Fig. 7 Comparison results of the algorithms with and without AE.

Table 4 Comparison results of the algorithms with different input methods.

Inputs	Alert (%)	Moderate fatigue (%)	Severe fatigue (%)	Average (%)
(X-landmark + Y-landmark +texture)	98.2	89.8	91.5	94.7
(X-landmark + Y-landmark) +texture	99.1	90.9	94.3	96.2

representative layer, and the bimodal DNN, which contains one joint layer with 600 units achieves the best performance. More units can lead to more parameters to be learned. However, a certain number of training samples cannot learn a neural network that contains too many parameters. By contrast, if the number of hidden units of each layer is too small, the 5520 input units are difficult to learn well. Our intention is to build a bimodal DNN with a certain number of parameters of joint representative layers to learn the texture and landmark modalities well and achieve high recognition accuracy in our dataset. As shown in Fig. 8, one joint layer with 600 units is adapted to learn the fusion representation of two modalities in this paper.

3.6.5 *Proposed algorithm versus other classifiers*

The fifth set of the experiment compares the performance between the proposed method and the other four classifiers, specifically decision tree (ID3), naïve Bayes, SVM, and k-nearest neighbor (KNN), which have been widely applied in many areas. The recognition results are illustrated in Table 5. As shown in the table, KNN outperforms the other classifiers except for the proposed algorithm, and Naïve Bayes has the poorest performance. Apparently, the recognition rate of "Alert" using the proposed method does not exceed SVM, but the average accuracy of the proposed algorithm is markedly superior to that of other classifiers.

3.6.6 *Comparison of the algorithms using difference frame rates*

In this experiment set, we perform a comparison of the methods using different frame rates. The experimental results are

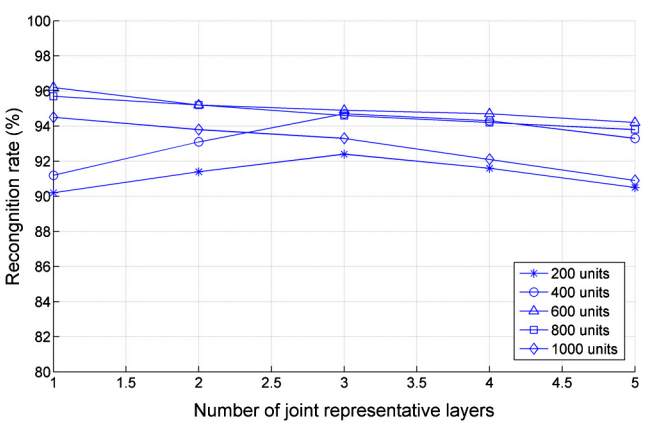

Fig. 8 Recognition results of the different parameters of the joint representative layers.

Table 5 Comparison the results using the proposed algorithm and other classifiers.

Inputs	Alert (%)	Moderate fatigue (%)	Severe fatigue (%)	Average (%)
ID3	68.0	59.1	67.9	66.1
Naïve Bayes	83.6	42.1	64.2	69.9
SVM	99.5	56.8	80.1	85.7
KNN	96.9	53.4	91.5	86.4
Proposed	99.1	90.9	94.3	96.2

shown in Fig. 9. As shown in the figure, the method using 12 frames per second (fps) achieves the best performance. It is because the facial behaviors of the subjects in our dataset are spontaneous (e.g., different facial expression and talking). When the frame rate extracted is small, some key dynamic information of facial behavior may be ignored. By contrast, a high frame rate will produce more input and hidden units, which can result in redundant information and more parameters to be learned. Our training data cannot afford a network that contains too many parameters, which can affect the performance of our method. From the experimental results, it can be concluded that the selection of the frame rate is based not only on the variations of facial behaviors but also on the number of training samples.

3.6.7 *Effect of different scenarios on recognition results*

In this experiment set, our algorithm is tested in two datasets obtained from an indoor environment (pretraining dataset) and an actual driving environment (training and testing dataset), respectively. The data preprocessing and validation strategy of the two experiments are the same. The recognition results are shown in Table 6. As shown in the table, the recognition rate obtained from indoor environment is higher than actual driving environment. It is clear that the conditions in actual driving environment are instable and have glares, shadows and dust, among others, which are the major challenges for neural classifier in driver fatigue detection systems. However, the gap between the recognition accuracies of two kinds of scenarios is not evidence because the landmark modality, which plays a major role in our model, is not susceptible to these interferences.

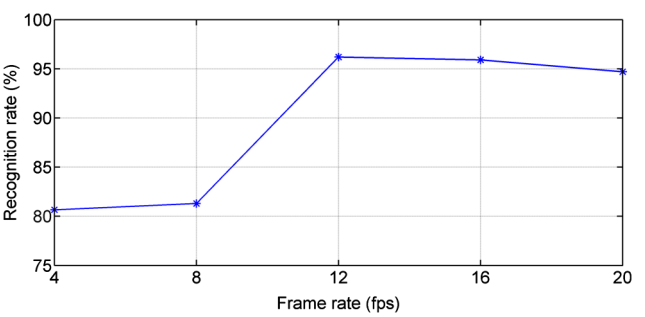

Fig. 9 Comparison of the methods using different frame rates.

Table 6 Comparison results of the algorithms tested in different scenarios.

Scenario	Alert (%)	Moderate fatigue (%)	Severe fatigue (%)	Average (%)
Indoor environment	99.8	93.3	97.6	97.7
Actual driving environment	99.1	90.9	94.3	96.2

4 Conclusions

This paper presents a human fatigue expression recognition method based on image-based dynamic multi-information and bimodal deep learning. The landmark and texture of the facial image sequences are combined and learned by bimodal DNN to boost the performance of fatigue recognition. The average recognition rate of the proposed methods is 96.2% on the dataset that we established. According to the results of the experiments, the dynamic landmark of the eyes and mouth and the texture of the eye region extracted from the image sequence can complement each other and reflect the driver's fatigue status correctly. To detect driver fatigue more accurately, large amounts of data are required to train the classifier to improve its generalization ability. Compared to shallow learning, the bimodal DNN can learn the landmark and texture well by unsupervised learning and perform more stably and robustly. Furthermore, DNN can avoid handcrafted features (e.g., Gabor, LBP, and HOG) to reduce computation complexity to entail low system cost. In practical application, the system that contains an ordinary HD camera and a image processing and recognition equipment with the core of a Texas Instruments' C6000 series digital signal processor can be use to implement our algorithm to detect driver fatigue in real time.

However, a limitation of the proposed method is that the occlusions and head rotations (yaw or pitch) can decrease the accuracy of fatigue recognition. This problem will be solved in our future work. A framework integrating other useful information (e.g., pupil movement) with the features (modalities) extracted in this paper could result in a good performance. This issue will also be studied in our future work. To prevent overfitting during the training process and to improve our method's generalization ability, we will extend the dataset to contain different races (e.g., American, European, and African) and a variety of possible situations such as illumination, occlusions, and head rotations, in the future. The ground truth of the driver fatigue was obtained from the subjective judgment of the researchers, which may make wrong judgments. This problem will be addressed by finding some objective approach in our future work.

Acknowledgments

This work was supported by the Open Foundation of State Key Laboratory of Automotive Simulation and Control (China, Grant No. 20121107). All participants involved in the experiment were volunteers who participated and understood the content and purpose of the study. The subjects shown in Figs. 2, 5, and 6 agreed to and authorized the publication of their facial images in this journal.

References

1. P. S. Rau, "Drowsy driver detection and warning system for commercial vehicle drivers: field operational test design, data analyses, and progress," in *Proc. of 19th Int. Conf. on Enhanced Safety of Vehicles*, 05-0192 (2005).
2. L. M. Bergasa et al., "Real-time system for monitoring driver vigilance," *IEEE Trans. Intell. Transp. Syst.* **7**(1), 63–77 (2006).
3. A. Sahayadhas, K. Sundaraj, and M. Murugappan, "Detecting driver drowsiness based on sensors: a review," *Sensors* **12**(12), 16937–16953 (2012).
4. M. Wang et al., "Drowsy behavior detection based on driving information," *Int. J. Automot. Technol.* **17**(1), 165–173 (2016).
5. Y. Takei and Y. Furukawa, "Estimate of driver's fatigue through steering motion," in *IEEE Int. Conf. on Systems, Man and Cybernetics*, pp. 1765–1770 (2005).
6. T. C. Chieh et al., "Driver fatigue detection using steering grip force," in *Proc. Student Conf. on Research and Development, (SCORED '03)*, pp. 45–48 (2003).
7. L. Barr et al., *A Review and Evaluation of Emerging Driver Fatigue Detection Measures and Technologies*, National Transportation Systems Center, US Department of Transportation, Cambridge, Washington (2005).
8. D. F. Dinges and R. Grace, *PERCLOS: A Valid Psychophysiological Measure of Alertness as Assessed by Psychomotor Vigilance*, US Department of Transportation, Federal Highway Administration, Washington, DC (1998).
9. C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Trans. Image Process.* **11**(4), 467–476 (2002).
10. W. Rong-Ben et al., "A monitoring method of driver fatigue behavior based on machine vision," in *Proc. IEEE Intelligent Vehicles Symp.*, pp. 110–113 (2003).
11. W. Dong and X. Wu, "Fatigue detection based on the distance of eyelid," in *Proc. of 2005 IEEE Int. Workshop on VLSI Design and Video Technology*, pp. 365–368 (2005).
12. X. Fan, B. Yin, and Y. Sun, "Nonintrusive driver fatigue detection," in *IEEE Int. Conf. on Networking, Sensing and Control (ICNSC '08)*, pp. 905–910 (2008).
13. K. W. Kim et al., "Segmentation method of eye region based on fuzzy logic system for classifying open and closed eyes," *Opt. Eng.* **54**(3), 033103 (2015).
14. M. M. Ibrahim, J. J. Soraghan, and L. Petropoulakis, "Eye-state analysis using an interdependence and adaptive scale mean shift (IASMS) algorithm," *Biomed. Signal Process. Control* **11**, 53–62 (2014).
15. F. Song et al., "Eyes closeness detection from still images with multi-scale histograms of principal oriented gradients," *Pattern Recognit.* **47**(9), 2825–2838 (2014).
16. B. Cyganek and S. Gruszczyński, "Hybrid computer vision system for drivers' eye recognition and fatigue monitoring," *Neurocomputing* **126**, 78–94 (2014).
17. W. Rongben et al., "Monitoring mouth movement for driver fatigue or distraction with one camera," in *Proc. 7th Int. IEEE Conf. on Intelligent Transportation Systems*, pp. 314–319 (2004).
18. T. Wang and P. Shi, "Yawning detection for determining driver drowsiness," in *Proc. of 2005 IEEE Int. Workshop on VLSI Design and Video Technology*, pp. 373–376 (2005).
19. L. Yufeng and W. Zengcai, "Detecting driver yawning in successive images," in *The 1st Int. Conf. on Bioinformatics and Biomedical Engineering (ICBBE '07)*, pp. 581–583 (2007).
20. S. J. Lee et al., "Real-time gaze estimator based on driver's head orientation for forward collision warning system," *IEEE Trans. Intell. Transp. Syst.* **12**(1), 254–267 (2011).
21. R. Oyini Mbouna, S. G. Kong, and M.-G. Chun, "Visual analysis of eye state and head pose for driver alertness monitoring," *IEEE Trans. Intell. Transp. Syst.* **14**(3), 1462–1469 (2013).
22. B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *Pattern Recognit.* **36**(1), 259–275 (2003).
23. R. A. Patil, V. Sahula, and A. S. Mandal, "Features classification using support vector machine for a facial expression recognition system," *J. Electron. Imaging* **21**(4), 043003 (2012).
24. Y. Li et al., "Measuring the intensity of spontaneous facial action units with dynamic Bayesian network," *Pattern Recognit.* **48**(11), 3417–3427 (2015).
25. B. Jiang and K. Jia, "Robust facial expression recognition algorithm based on local metric learning," *J. Electron. Imaging* **25**(1), 013022 (2016).
26. X. Fan and T. Tjahjadi, "A spatial-temporal framework based on histogram of gradients and optical flow for facial expression recognition in video sequences," *Pattern Recognit.* **48**(11), 3407–3416 (2015).
27. H. Sadeghi and A. A. Raie, "Suitable models for face geometry normalization in facial expression recognition," *J. Electron. Imaging* **24**(1), 013005 (2015).
28. H. Gu and Q. Ji, "An automated face reader for fatigue detection," in *Proc. Sixth IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pp. 111–116 (2004).

29. X. Fan et al., "Gabor-based dynamic representation for human fatigue monitoring in facial image sequences," *Pattern Recognit. Lett.* **31**(3), 234–243 (2010).
30. C. Zhao et al., "Driver's fatigue expressions recognition by combined features from pyramid histogram of oriented gradient and contourlet transform with random subspace ensembles," *Intell. Transp. Syst.* **7**(1), 36–45 (2013).
31. Y. Zhang and C. Hua, "Driver fatigue recognition based on facial expression analysis using local binary patterns," *Optik-Int. J. Light Electron Opt.* **126**(23), 4501–4505 (2015).
32. C. Zhao et al., "Classification of driver fatigue expressions by combined curvelet features and Gabor features, and random subspace ensembles of support vector machines," *J. Intell. Fuzzy Syst.* **26**(1), 91–100 (2014).
33. P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vision* **57**(2), 137–154 (2004).
34. A. Asthana et al., "Incremental face alignment in the wild," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1859–1866 (2014).
35. G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science* **313**(5786), 504–507 (2006).
36. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature* **521**(7553), 436–444 (2015).
37. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Advances in Neural Information Processing Systems*, pp. 1097–1105 (2012).
38. G. Hinton et al., "Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups," *IEEE Signal Process. Mag.* **29**(6), 82–97 (2012).
39. J. Ma et al., "Deep neural nets as a method for quantitative structure-activity relationships," *J. Chem. Inf. Model.* **55**(2), 263–274 (2015).
40. T. Ciodaro et al., "Online particle detection with neural networks based on topological calorimetry information," *J. Phys. Conf. Ser.* **368**(1), 012030 (2012).
41. Y. Bengio et al., "Greedy layer-wise training of deep networks," in *Proc. Advances in Neural Information Processing Systems*, Vol. 19, p. 153 (2007).
42. I. Sutskever, J. Martens, and G. E. Hinton, "Generating text with recurrent neural networks," in *Proc. of the 28th Int. Conf. on Machine Learning (ICML-11)*, pp. 1017–1024 (2011).
43. D. R. Williams and G. E. Hinton, "Learning representations by back-propagating errors," *Nature* **323**, 533–536 (1986).
44. J. Ngiam et al., "Multimodal deep learning," in *Proc. of the 28th Int. Conf. on Machine Learning (ICML '11)*, pp. 689–696 (2011).
45. N. Srivastava and R. Salakhutdinov, "Learning representations for multimodal data with deep belief nets," in *Int. Conf. on Machine Learning Workshop* (2012).
46. N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," in *Advances in Neural Information Processing Systems*, pp. 2222–2230 (2012).
47. W. Zhang et al., "Multimodal learning for facial expression recognition," *Pattern Recognit.* **48**(10), 3191–3202 (2015).
48. H. Jung et al., "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proc. of the IEEE Int. Conf. on Computer Vision*, pp. 2983–2991 (2015).
49. A. Krizhevsky and G. Hinton, *Learning Multiple Layers of Features from Tiny Images*, Technical Report, University of Toronto (2009).

Lei Zhao received his BS and MS degrees from Shandong University of Science and Technology, China, in 2011 and 2014, respectively. He is currently pursuing his PhD at Shandong University, China. His research interests include image processing, pattern recognition, and vehicle active safety.

Zengcai Wang received his BS and MS degrees from Shandong University of Science and Technology, China, in 1983 and 1988, respectively, and his PhD from China University of Mining and Technology, China, in 1999. He is currently a professor and a doctoral supervisor of vehicle engineering at Shandong University. His research interests include image processing and computer vision, pattern recognition, vehicle active safety, and mechanical and electrical hydraulic control system.

Xiaojin Wang received his BS degree from Shandong University in 2014. He is currently pursuing his MS degree at Shandong University, China. His research interests include computer vision, pattern recognition, and vehicle active safety.

Yazhou Qi received his BS degree from Hubei University of Automotive Techology in 2014. He is currently pursuing his MS degree at Shandong University, China. His research interests include computer vision, pattern recognition, and vehicle active safety.

Qing Liu received his BS and MS degrees from Shandong University of Science and Technology, China, in 2012 and 2015, respectively. He is currently pursuing his PhD at Shandong University, China. His research interests include computer vision, pattern recognition, and vehicle active safety.

Guoxin Zhang received his BS degree from Shandong University in 2014. He is currently pursuing his MS degree at Shandong University, China. His research interests include computer vision, pattern recognition, and vehicle active safety.