Driver Facial Fatigue Behavior Recognition Using Local Ensemble Convolutional Neural Network and Encoding Vector

Lei, Zhao[1,*] and Guoxin Zhang[2]

[1] *School of Mechanical and Electronic Engineering, Shandong Jianzhu University, No.1000 of Fengming Road, Jinan, China*
[2] *Department of Mechanical Engineering, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong*
[*] *Corresponding author. E-mail: leizhao1219@sdjzu.edu.cn*

Received: Aug. 07, 2019, Accepted: Mar. 26, 2020

A novel approach for the detecting dynamic facial fatigue behavior of drivers is proposed in this work. First, the eye and mouth regions are extracted from a frame in a video to classify the static facial fatigue state by using a local ensemble convolutional neural network (LECNN) model. A transfer learning strategy is used to pretrain the LECNN model for improved recognition accuracy. Second, the classification results for the frames in the video are encoded and connected to construct an encoding vector that represents the facial state in the video. Finally, k-nearest neighbors classifier is adopted to classify the constructed vector and obtain the recognition result of facial fatigue. The proposed system is tested on a human fatigue dataset of images that are captured in different environments. Results demonstrate the stable and robust performance of the proposed method with approximately 96.76% accuracy.

Keywords: Driver fatigue recognition; Local ensemble convolutional neural networks; Transfer learning

http://dx.doi.org/10.6180/jase.202009_23(3).0001

1. Introduction

In contrast with drunk driving, overspeeding, and other major types of risky driving behaviors, driver fatigue is difficult to detect and prevent, thereby compromising the security and safety of drivers, passengers, and pedestrians. Therefore, studying the effective methods for detecting human drowsiness is crucial in improving transport safety. Eye-based human fatigue detection methods have been widely used. In [1], an eye state classification method that used the feature-level fusion and user-specific classification of both eyes, was introduced to improve the accuracy of driver fatigue recognition. Cyganek et al. [2] proposed a detection method that determined the states of driver fatigue via eye recognition using two cameras. Dasgupta [3] presented a robust real-time embedded platform to monitor the attention loss of drivers under day and night driving conditions. Facial expression recognition, which is a significant and challenging problem in the fields of pattern recognition and computer vision, has recently attracted considerable attention. The method in [4] employs Gabor wavelet and curvelet transforms to extract features from static facial images and uses the random subspace ensembles of support vector machines (SVMs) with a polynomial kernel to recognize the three categories of facial fatigue expression. Zhao et al. [5] proposed a high-order local multiresolution derivative pattern to extract the discriminative information of the contours of the mouth, eyes, and eyebrows from the sparse representation of fatigue expression images and adopted the kernel SVM classifier to detect human fatigue. In recent years, many driver facial fatigue behavior recognition methods based on deep neural network have been proposed [6]. Jabbar [7] fused facial images and feature locations and used a real-time deep learning model to identify driver fatigue status.

In this work, a local ensemble convolutional neural network (LECNN) model is proposed to detect driver facial fatigue in state images. A transfer learning strategy is applied to pretrain the convolutional neural networks (CNNs) in the LECNN model by using different datasets to reduce the detrimental effect of limited data on recognition accuracy. The classification results for the frames in two-second videos are encoded and connected to construct an encoding vector that represents the statistical information of the facial state in the video. Subsequently, a k-nearest neighbors (k-NN) classifier is used to classify the vector and obtain facial fatigue recognition results.

2. Theory and formula

The proposed method comprises face and facial point detection, LECNN development, transfer learning, encoding vector construction, and dynamic facial fatigue classification. The exhaustive illustration of the proposed method is presented in Fig. 1.

2.1. Face and facial point detection

A real-time Viola–Jones face detector [8] is utilized to detect the facial region when a video is inputted to a human fatigue recognition system. The average rate of frames at which face detection occurs is not less than 50 frames per second in our platform. Then, an improved version of a constrained local model [9] is adopted to locate and track facial landmarks.

2.2. LECNN

CNNs are feed-forward neural networks [10] that comprise convolutional, pooling, and fully connected neural network layers. Convolution is mathematically represented as follows:

$$v_{ij}^{xy} = f\left(b_{ij} + \sum_{k=0}^{M-1}\sum_{p=0}^{p_i-1}\sum_{q=0}^{p_i-1} w_{ijk}^{pq} v_{(i-1)k}^{(x+p)(y+q)}\right) \tag{1}$$

where, v_{ij}^{xy} is the value at (x, y) in the jth map in the ith layer; f is the activate function; b_{ij} is the bias; M is the number of maps; w_{ijk}^{pq} is the weight parameter at location (p, q) connected to the kth map; and P_i and Q_i are the values of the kernel height and width, respectively. Pooling layers are adopted to reduce the map size, thereby avoiding over-fitting and increasing the position invariance of the kernels. Fully connected layers are similar to general neural networks, and their neurons are fully connected to the previous layer. In this study, eye and mouth patches are cropped and fed into the LECNN model. The architecture of the proposed LECNN model is presented in Fig. 2. The LECNN model comprises three identical CNNs. The model receives three 24×24 grayscale images, which represent the eyes and the mouth as inputs. Finally, the hidden layers of these CNNs are concatenated and connected to the output layers as follows:

$$A = [a_1, a_2, a_3] \tag{2}$$

$$p(y = i|A) = \sigma\left(\sum_{k=0}^{K} A_k W_{k,i}\right) \tag{3}$$

where A is the concatenated layer; a_1, a_2, and a_3 are the fully connected hidden layers of the three CNNs; $p(y = i|A)$ is the predicted probability for the ith class given hidden vector A; σ is the SoftMax function; and $W = [W_{1i}, W_{2i}, \ldots, W_{Ki}]$ is the parameter set of the output layers. The inference can be obtained by the following function:

$$\hat{I} = \underset{i}{argmax} P \tag{4}$$

where P is the vector that contains all the predicted probabilities for the classes of the network.

2.3. Transfer learning

The lack of training data can render the proposed network prone to over-fitting. Accordingly, a suitable transfer learning method for the LECNN is developed. In the LECNN, the structures of the three CNNs are identical. A CNN model is trained using the pretraining dataset, and the learned features are subsequently transferred to the three CNNs of the LECNN. After pretraining, the pretrained LECNN is fine-tuned on the constructed dataset. Five facial expression recognition datasets (i.e., CK+ [11], MMI [12], FERA [13], FER2013 [14] and JAFFE [15]) and an image classification dataset (CIFR-10 [16]) are adopted study to pretrain the LECNN.

2.4. Encoding vector

The classification results for the frames in the video are encoded and connected to construct an encoding vector that represents the time series information of the facial static states in the video. The recognition results for the extracted frames are numbered on the basis of Formula (4) and concatenated into a vector as follows:

$$V = [\hat{I}_1, \hat{I}_2, ..., \hat{I}_i, ...\hat{I}_n] \tag{5}$$

where V is the encoding vector of the video, $\hat{I}_i$ is the numbered recognition result of the ith frame in the video, and n is the number of frames extracted from the two-second video. Finally, a k-NN classifier is applied to classify the vector and obtain the recognition result. The k-NN classifier algorithms is a type of instance-based learning model. The k-NN is to predict the target label by finding the nearest neighbor class. The closest class will be identified using the distance measures like Euclidean distance. KNN can be used for both classification and regression predictive problems. However, it is more widely used in classification problems in many fields.

3. Experimental Dataset

The actual driving environment of driver facial fatigue behavior is captured to evaluate the proposed method. An

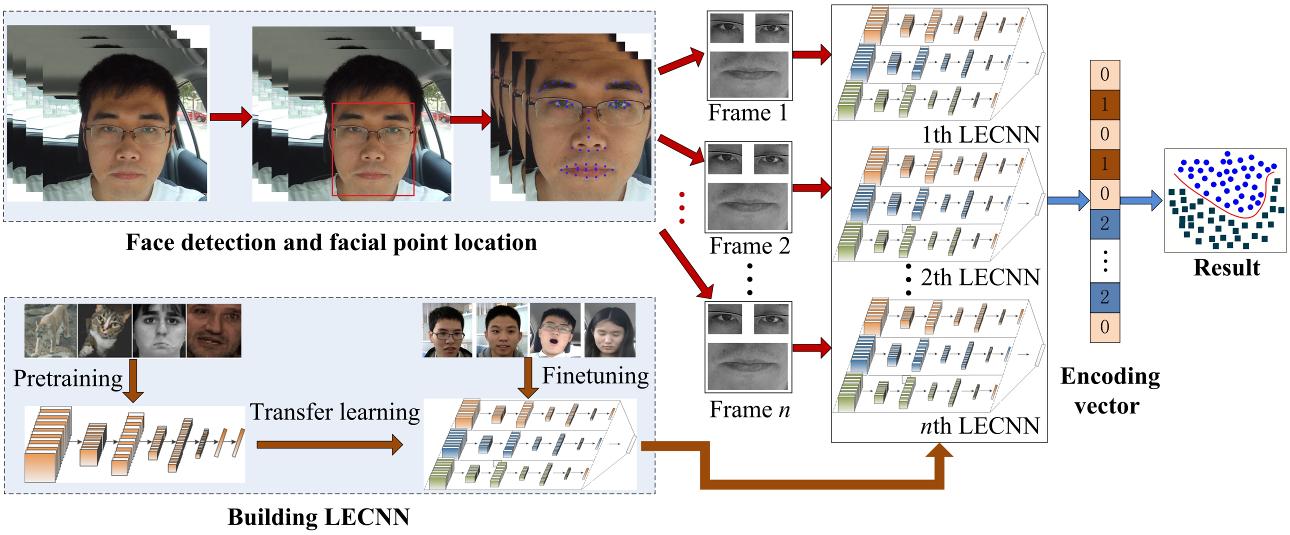

Fig. 1. Overall structure of the proposed framework.

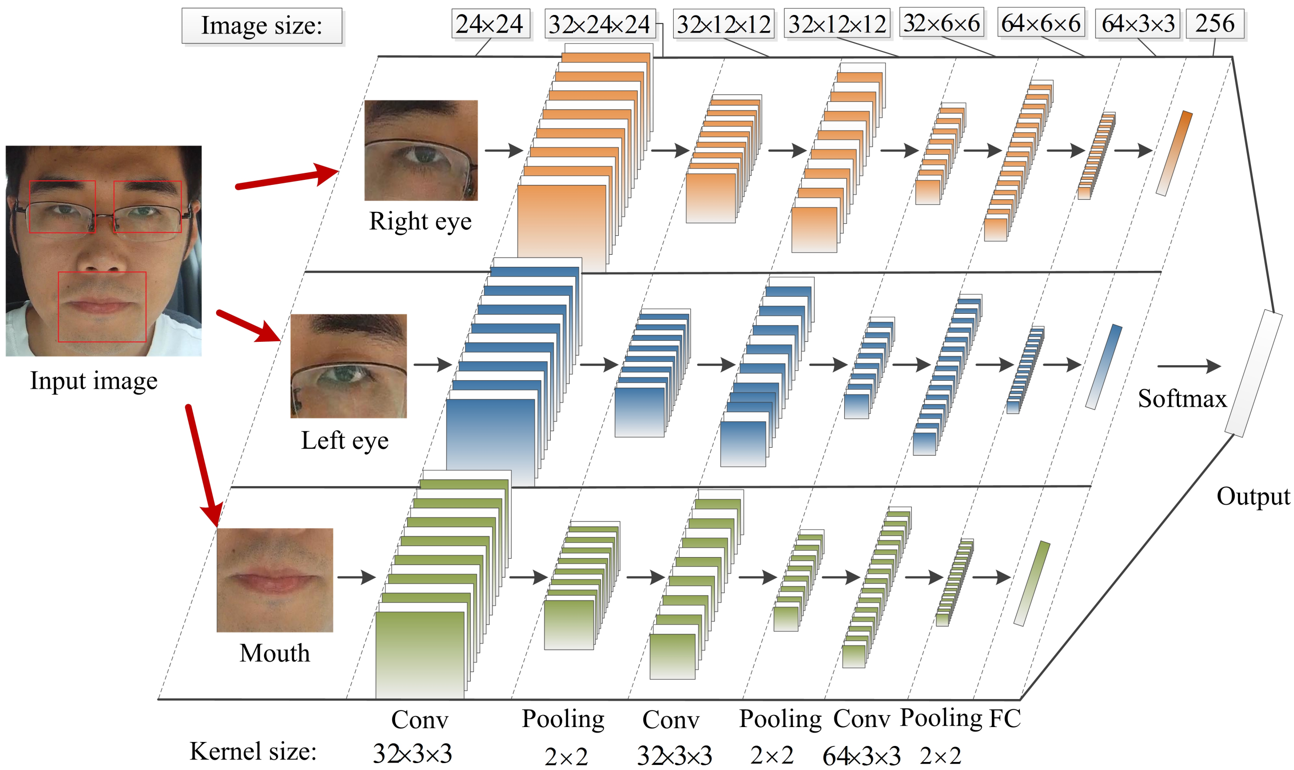

Fig. 2. Architecture of the proposed LECNN.

AHD130W-A381 near infrared camera is installed in front of the subject's seat on the right side of the driver to capture the subjects' face considering the dangers of driver fatigue and the invariance of fatigue expressions on different occasions. The videos of 30 subjects are captured in driving environments, and the ages of the subjects ranged from 20–55 years. Females account for approximately 24% of the subjects. Video captures include day, evening (dim light), and night scenes. The facial behavior of each subject is spontaneous. Finally, all 586 videos are converted into AVI format and compressed by MPEG-4, and the video rate is 24 fps. The length of each video is 2 s (48 frames). In addition, 22870 images of the 586 videos are extracted to train and evaluate the proposed LECNN model. A subject-based 10-fold cross-validation method is adopted in the experiments. The classification scheme, namely, alert, moderate fatigue, and severe fatigue, is adopted in our dataset. Alert indicates that the driver drives the car in alert state. Moderate fatigue indicates that the driver is in fatigue state, characterized by yawning and frequent blinking. Severe fatigue is characterized by sleeping and loss of ability to control the car. The details of the datasets are shown in Table 1.

Table 1. Details of the dataset.

Dataset	Subjects	Alert	Moderate fatigue	Severe fatigue	Total
Video	30	324	136	126	586
Image	30	13410	4503	4957	22870

4. Result and discussions

First, the LECNN model is compared with the LBP, Gabor, and HOG feature extraction methods. The accuracy of each classifier is shown in Table 2. The experimental results suggest that the overall recognition accuracy of the proposed LECNN model is higher than those of other methods regardless of the classifier used to classify the extracted features. The recognition results also show that the LECNN model with transfer learning strategy (LECNNT) outperforms the model without transfer learning (LECNNNT). Subsequently, the output state of each frame is used to construct the encoding vector. A k-NN is used to classify the vector to obtain the driver fatigue state.

Then, the proposed facial fatigue behavior recognition method is compared with other methods [17, 18]. The experiment results are shown in Table 3. SLBP corresponds to spatial-temporal motion local binary pattern [18]. The experimental result shows that the proposed method outperforms two other state-of-the-art methods in driver facial fatigue recognition. Temporal resolution is defined as the number of frames extracted from a video for classification. The recognition accuracy rates of state-of-the-art methods for different temporal resolutions are analyzed and compared in this section. The experimental results are shown in Fig. 3. The classification capabilities of the methods are considerably improved when the extracted number of frames is increased. The proposed method outperforms the other methods in all temporal resolutions.

5. Conclusion

This paper presents a driver facial fatigue behavior recognition method based on a LECNN and an encoding vector.

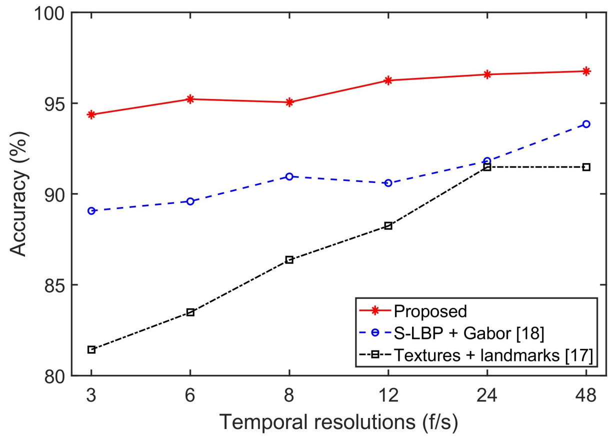

Fig. 3. Recognition results for different temporal resolutions.

The eyes and mouth images are fed into the LECNN, which integrates three CNNs, to classify the facial fatigue behavior in static frames. A transfer learning strategy is applied to pretrain the CNNs in the LECNN by using different datasets to improve the recognition accuracy. The classification results for the frames in the video are encoded and connected to form an encoding vector that represents the statistical information of the facial state in the video. A k-NN is used to classify the vector. The experimental results show that the LECNN can accurately recognize static facial fatigue. The average recognition rate of the proposed method on the constructed dataset is 96.76%.

In the future, we will expand our dataset to include more complex environments (dim and night environments), people (e.g., eye disabled people), and races (e.g., European, American, and African) to obtain more reliable recognition results.

Acknowledgments

This work was supported by the Doctoral fundation of shandong jianzhu university (X18039Z) and the Natural Science Foundation of Shandong Province (ZR2018MEE015).

Table 2. Comparison of the results of LECNN and other methods.

Feature	Classifier	Alert	Moderate fatigue	Severe fatigue	Average
LBP	*k*-NN	89.78	80.08	85.88	87.02
	RF	97.78	66.87	84.65	88.85
	SVM	96.59	83.90	88.38	92.31
HOG	*k*-NN	90.58	66.40	80.90	83.73
	RF	97.68	76.75	83.80	90.55
	SVM	95.71	85.01	89.43	92.24
Gabor	*k*-NN	93.62	76.11	89.91	89.37
	RF	96.91	88.67	85.96	92.92
	SVM	96.05	86.12	92.31	93.28
Image LECNN$_{NT}$	96.36	95.00	90.84	94.90	
	LECNN$_{T}$	97.36	95.40	93.44	96.13

Table 3. Comparison results of the proposed method and four state-of-the-art methods.

Methods	Alert	Moderate fatigue	Severe fatigue	Average
Texture + landmarks [17]	94.44	91.18	84.92	91.64
SLBP + Gabor [18]	96.91	90.44	89.68	93.85
proposed	98.46	96.32	92.86	96.76

References

[1] Jaeik Jo, Sung Joo Lee, Kang Ryoung Park, Ig Jae Kim, and Jaihie Kim. Detecting driver drowsiness using feature-level fusion and user-specific classification. *Expert Systems with Applications*, 41(4 PART 1):1139–1152, jan 2014.

[2] Bogusław Cyganek and Sławomir Gruszczyński. Hybrid computer vision system for drivers' eye recognition and fatigue monitoring. *Neurocomputing*, 126:78–94, feb 2014.

[3] Anirban Dasgupta, Anjith George, S. L. Happy, and Aurobinda Routray. A vision-based system for monitoring the loss of attention in automotive drivers. *IEEE Transactions on Intelligent Transportation Systems*, 14(4):1825–1838, dec 2013.

[4] Chihang Zhao, Jie Lian, Qian Dang, and Can Tong. Classification of driver fatigue expressions by combined curvelet features and gabor features, and random subspace ensembles of support vector machines. *Journal of Intelligent and Fuzzy Systems*, 26(1):91–100, 2014.

[5] Chihang Zhao, Yunsheng Zhang, Xiaozheng Zhang, and Jie He. Recognition of driver's fatigue expression using Local Multiresolution Derivative Pattern. *Journal of Intelligent and Fuzzy Systems*, 30(1):547–560, 2016.

[6] Changyuan Wang, Ting Yan, and Hongbo Jia. Spatial-Temporal Feature Representation Learning for Facial Fatigue Detection. *International Journal of Pattern Recognition and Artificial Intelligence*, 32(12):1856018, dec 2018.

[7] Rateb Jabbar, Khalifa Al-Khalifa, Mohamed Kharbeche, Wael Alhajyaseen, Mohsen Jafari, and Shan Jiang. Real-time Driver Drowsiness Detection for Android Application Using Deep Neural Networks Techniques. In *Procedia Computer Science*, volume 130, pages 400–407, 2018.

[8] Paul Viola and Michael J. Jones. Robust Real-Time Face Detection. *International Journal of Computer Vision*, 57(2):137–154, may 2004.

[9] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. Incremental face alignment in the wild. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1859–1866, 2014.

[10] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2323, 1998.

[11] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010*, pages 94–101, 2010.

[12] Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. Web-based database for facial expression analysis. In *2005 IEEE International Conference on Mul-*

timedia and Expo, pages 317–321, Amsterdam, Netherlands, 2005.

[13] Tanja Banziger and Klaus R Scherer. Introducing the Geneva Multimodal Emotion Portrayal (GEMEP) Corpus. *A Blueprint for Affective Computing: A sourcebook and manual*, pages 271–294, 2010.

[14] Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang, and Yoshua Bengio. Challenges in representation learning: A report on three machine learning contests. In *International Conference on neural Information Processing*, pages 117–124, Daegu, korea, 2013.

[15] Michael J. Lyons. Automatic classification of single facial images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(12):1357–1362, dec 1999.

[16] Alex Krizhevsky and G Hinton. *Learning multiple layers of features from tiny images*. Toronto, ON, Canada, 2009.

[17] Lei Zhao, Zengcai Wang, and Guoxin Zhang. Facial Expression Recognition from Video Sequences Based on Spatial-Temporal Motion Local Binary Pattern and Gabor Multiorientation Fusion Histogram. *Mathematical Problems in Engineering*, 2017, 2017.

[18] Lei Zhao, Zengcai Wang, Xiaojin Wang, Yazhou Qi, Qing Liu, and Guoxin Zhang. Human fatigue expression recognition through image-based dynamic multi-information and bimodal deep learning. *Journal of Electronic Imaging*, 25(5):053024, oct 2016.