Postprint: Research on a Fatigue Detection Method Based on Multi-scale Pooling Convolutional Neural Networks

Authors: Gu Wanghuan, Zhu Yu, Chen Xudong, Zheng Bingbing, He Linfei

Date: 2018-08-13T00:00:00+00:00

Abstract

Driver fatigue constitutes one of the primary causes of traffic accidents, and detecting driver fatigue through visual sensors that capture facial images and perform visual feature analysis holds practical significance. To address the challenge of fatigue detection via visual feature analysis, we design a cascaded deep learning-based detection system architecture and propose a fatigue state detection model employing convolutional neural networks with multi-scale pooling. Initially, face detection is conducted using the deep learning model MTCNN to extract eye and mouth regions. To tackle the representation and recognition of eye and mouth states, we propose a ResNet-based multi-scale pooling model (MSP) for training on eye and mouth states. During real-time detection, the eye and mouth regions are processed through the trained convolutional neural network model for state recognition, and driver fatigue is ultimately determined based on PERCLOS and the proposed frequency of mouth opening (FOM). Experimental results demonstrate that the algorithm achieves high detection accuracy, satisfies real-time requirements, and exhibits robust performance in complex environments.

Full Text

Preamble

Abstract: Fatigue driving is one of the main causes of traffic accidents, particularly for drivers of large vehicles like buses and heavy trucks. For designing driver fatigue monitoring systems, visual features based techniques are among the most effective approaches. This paper proposes a hierarchical convolutional neural network model with multi-scale pooling for vision-based fatigue detection systems. The first step involves face detection and extraction of eye and mouth regions using the deep learning model MTCNN. To address the problem of characterization and recognition of eye and mouth regions, this paper proposes a multi-scale pooling model (MSP) based on ResNet to train different states of eyes and mouth. During real-time detection, the states of eyes and mouth are recognized by the pre-trained convolutional neural network model. Finally, fatigue is detected through PERCLOS and the frequency of open mouth (FOM). Experimental results show that the proposed algorithm has high detection accuracy, real-time performance, and high robustness to complex environments.

Key words: visual feature analysis; multi-scale pooling; convolutional neural network; fatigue detection; face detection

Introduction

Based on the analysis above, fatigue detection algorithms using driver facial behavior features generally consist of two steps: (a) face detection, and (b) fatigue judgment based on facial behavior states. According to a survey by the AAA Foundation for Traffic Safety [1], 16%-21% of traffic accidents are caused by driver fatigue, particularly severe for long-haul trucks and buses. Therefore, driver fatigue detection technology holds significant research importance for accident prevention.

When entering a fatigue state, humans typically exhibit various biological changes such as yawning, prolonged eye closure, and changes in brain waves and heart waves. This paper focuses fatigue detection research primarily on analyzing driver facial behavior features. Numerous studies have investigated face detection algorithms. Reference [2] presents Viola' s classic work, which first introduced the integral image concept combined with the Adaboost algorithm for face detection, achieving real-time monitoring while ensuring detection accuracy. Yong Du et al. [3] proposed a method for fatigue judgment based on driver eye state changes using frame difference methods with color information for face detection. In 2012, convolutional neural networks (CNN) [4] made a breakthrough in the ImageNet competition, making deep learning (DL) the most popular approach for image and video processing in recent years, achieving tremendous success in image retrieval [5], human action recognition [6], and scene analysis [7]. Deep learning methods have also proven remarkably effective for face detection and recognition [8]. Taigman et al. [9] proposed the DeepFace model in 2014, successfully applying deep learning to face detection and recognition, solving the common robustness issues in traditional methods and significantly improving accuracy. In 2016, Zhang et al. [10] proposed a multi-task convolutional neural network model (MTCNN) that trains face detection across three network layers. This model can quickly and accurately detect multiple faces with good robustness for profile faces, achieving 95% accuracy on the FDDB, WIDERFACE, and AFLW datasets.

Methods for fatigue judgment based on facial states primarily analyze eye openness and yawning behavior. Wierwille et al. [11] established the PERCLOS theoretical model, defining PERCLOS as the proportion of time within a unit period when eyelid closure reaches 70% or 80%. PERCLOS has become a highly effective fatigue assessment metric. Chu et al. [12] studied mouth shapes, feeding mouth data into a BP network to classify three different mental states. Deep learning-based fatigue detection has become a hot research topic in recent years. Zhao Xuepeng et al. [13] used a cascaded network to locate and detect eye regions for fatigue judgment. Geng Lei et al. [14] first employed AdaBoost and KCF (kernelized correlation filters) for face detection and tracking, then used classic network structures to detect eye and mouth states. Vision-based fatigue detection has been studied for many years. However, due to significant interference from environmental lighting and facial expressions in real-world scenarios, there remains room for improvement in face detection accuracy and facial behavior feature detection. To address this issue, this paper proposes a cascaded deep learning structure and a fatigue state detection model based on multi-scale pooling convolutional neural networks (MSP-net). The system detects faces using a pre-trained MTCNN model, extracts eye and mouth data, and feeds these into a pre-trained MSP-Net CNN to determine eye openness and mouth closure. Finally, fatigue is jointly determined based on PERCLOS and mouth opening frequency (FOM). The overall process is shown in [Figure 1: see original paper].

1.1 Data Collection

The network training data for eye and mouth image state determination was independently collected by the authors during practical work, including actual driver images and assistance from 21 volunteers. Considering nighttime driving scenarios, the collection equipment included both regular cameras and infrared cameras. During image and video acquisition, various complex environmental conditions encountered during actual driving were comprehensively considered. The collected image data included eyes open/closed, mouth open/closed, no glasses, wearing glasses, frontal and profile views. After collection, screening was performed to remove noisy samples. The final filtered samples were categorized into: regular camera eyes open, infrared camera eyes open, regular camera eyes closed, infrared camera eyes closed, regular camera mouth open, infrared camera mouth open, regular camera mouth closed, and infrared camera mouth closed. The dataset comprised 36,764 eye samples and 15,185 mouth samples. Sample examples are shown in [Figure 2: see original paper].

1.2 Data Preprocessing

Since infrared camera samples generally have low brightness, direct network training yields poor results. Therefore, sample preprocessing is necessary. This paper employs histogram equalization [16] to enhance local contrast. The main steps involve calculating the cumulative probability of the original grayscale levels and mapping the original grayscale values to new values based on this relationship. For grayscale images with values 0-255, the grayscale mapping relationship is:

$$p_k = 255 \times \sum_{i=0}^{k} \frac{n_i}{n}, \quad k = 0, 1, 2, \dots, L-1$$

where p_k represents the mapped grayscale value; k denotes the k-th grayscale level; L is the total number of grayscale levels; n is the total number of pixels; and n_i is the number of pixels at the i-th grayscale level. The resulting p_k is then rounded. [Figure 3: see original paper] shows the comparison before and after histogram equalization.

2 Face Detection

Traditional face detection algorithms such as AdaBoost or frame difference methods exhibit poor robustness in complex environments, where deep learning models have significant advantages. This paper adopts Zhang et al.' s MTCNN model for face detection. The MTCNN network structure, shown in [Figure 4: see original paper], consists of three main layers: P-NET, R-NET, and O-NET.

P-NET: First constructs an image pyramid, then uses a fully convolutional network (FCN) [17] to obtain candidate facial windows and bounding box regression vectors. The bounding box regression calibrates the candidate windows, followed by non-maximum suppression (NMS) [18] to merge highly overlapping candidate regions.

R-NET: Takes candidate regions from P-NET for further screening of false positives and calibration, again using NMS to merge candidate regions.

O-NET: Similar to R-NET but performs more refined supervision on candidate regions, finally outputting 5 additional keypoints.

MTCNN demonstrates excellent robustness, accurately detecting faces even at certain rotation angles. By detecting faces with MTCNN and using the returned 5 keypoints, eye and mouth regions can be successfully located. As shown in [Figure 5: see original paper], the left image shows detection results under a regular camera, while the right shows infrared camera results.

3 Multi-Scale Pooling Convolutional Neural Network MSP-Net

After extracting eye and mouth regions, state detection is required. This paper proposes a multi-scale pooling convolutional neural network model (MSP-Net). Two separate MSP-Net models are trained for eyes and mouth respectively.

3.1 Model Design

The multi-scale pooling convolutional neural network model (MSP-net) improves upon the ResNet [19] structure, retaining the residual concept while modifying the pooling layer by replacing max pooling with the proposed multi-scale pooling (MSP) to improve recognition performance for images captured at different resolutions. The MSP structure is shown in [Figure 6: see original paper].

MSP module steps: (a) The feature maps from the previous layer undergo two max pooling operations to obtain feature maps with side lengths 1/4 of the original; (b) The original feature maps are scaled down by half, and the resulting new feature maps undergo one max pooling to obtain another set of feature maps 1/4 the original size; (c) The original feature maps are scaled down to 1/4 of their original side length; (d) The three output feature maps are concatenated and fed into subsequent deep learning networks.

The multi-scale pooling concept originates from spatial pyramid pooling models [20]. Compared to spatial pyramid models, multi-scale pooling offers more flexible placement, allowing multiple uses at the beginning, middle, or end of networks.

The MSP-net structure designed in this paper is shown in [Figure 7: see original paper]. Eyes and mouth are trained identically; here we use eyes as an example. In MSP-net, the input is a 48$\times$48 grayscale image. After one convolution and MSP operation, it outputs $12\times12\times4$ feature maps. The feature maps are then flattened into a one-dimensional vector and fed into a fully connected layer with input length 1,728, a hidden layer of length 1,000, and finally softmax output with four categories: regular camera eyes open, infrared camera eyes open, regular camera eyes closed, and infrared camera eyes closed.

To verify MSP-net effectiveness, this paper conducts experiments using classic AlexNet [4] and ResNet structures as comparison networks. Section 6.2 shows comparative results for the three networks.

3.2 Loss Function and Optimization Method

The network uses softmax classification with four categories. The softmax function is defined as:

$$p_j = \frac{\exp(y_j)}{\sum_{k=0}^{3} \exp(y_k)}, \quad j = 0, 1, 2, 3$$

where p_j represents the probability of class j; $y'_j = \sum h_i \cdot w_{i,j} + b_j$ denotes the output of the final fully connected layer, with h_i being the previous layer output, and $w_{i,j}$ and b_j the weights and biases of the final layer.

The loss function is defined as cross-entropy:

$$L_m = -\sum_{j=0}^{3} 1\{y = j\} \log p_j$$

where L_m is the cross-entropy for the m-th sample; $1\{y = j\}$ is the indicator function equal to 1 when $y = j$ and 0 otherwise. Equation (3) shows the loss for a single sample; for M training samples, the loss function averages over all samples:

$$L = \frac{1}{M} \sum_{m=1}^{M} L_m$$

The optimization method uses Adaptive Moment Estimation (Adam) [21].

4 Fatigue State Detection

4.1 PERCLOS

PERCLOS represents the ratio between the number of eye-closure frames and total frames within a unit time [11]. The calculation formula is:

$$f_{per} = \frac{n}{N} \times 100$$

where n is the number of eye-closure frames and N is the total number of frames in the unit time. PERCLOS effectively quantifies eye closure duration. When PERCLOS reaches a certain threshold (literature [11] suggests 0.15), it indicates prolonged eye closure and preliminary fatigue.

4.2 Mouth Opening Frequency (FOM)

Similar to PERCLOS, the frequency of open mouth (FOM) represents the ratio between the number of mouth-open frames and total frames within a unit time. The calculation formula is:

$$f_{fom} = \frac{n}{N} \times 100$$

Like PERCLOS, larger values indicate greater fatigue. Final fatigue detection requires joint consideration of both metrics.

4.3 Fatigue State Detection

After completing pre-training of all deep networks, fatigue state thresholds are determined based on PERCLOS and FOM for real-time detection. The specific steps are: the camera captures driver video, MTCNN extracts faces and 5 key-points per frame to locate eye and mouth regions, MSP-net detects the state of extracted eye and mouth regions per frame, results are saved in a fixed-length queue, and the algorithm monitors queue value changes in real-time. When the distribution of all values in the queue reaches the threshold fatigue state, the alarm mechanism activates to alert the driver.

5 Experiments and Analysis

5.1 Experimental Platform and Dataset

All sample collection, network training, and result testing were implemented using Python 3.6 and TensorFlow 1.2. Cameras included regular cameras and 920 nm infrared cameras. The processor was a 3.20 GHz CPU with 4 GB RAM. The eye and mouth database (EMD) was collected using both camera types from 21 volunteers, comprising 36,764 eye samples and 15,185 mouth samples.

To further validate MSP-Net detection effectiveness, experiments were conducted on the public ZJU Eyeblink Database [22] ([Figure 8: see original paper]). This database contains 80 video clips from 20 individuals: each person contributed 4 clips (no glasses frontal view, thin-rim glasses frontal view, black-rim glasses frontal view, and no glasses upward view). The dataset is divided into open-eye and closed-eye categories, with 7,000 open-eye samples (5,770 training, 1,230 testing) and 1,984 closed-eye samples (1,574 training, 410 testing). Each image is 24$×$24 pixels.

5.2.1 Eye and Mouth State Detection

The self-built EMD dataset contains 36,764 eye samples and 15,185 mouth samples, with 95% for training and 5% for testing. Detailed categories and sample counts are shown in . Before training, operations including grayscale conversion, normalization, and histogram equalization were performed, with 48$×$48 grayscale images input to the network. To verify MSP-Net efficiency for eye and mouth state detection, classic AlexNet and ResNet were also trained and tested for comparison. Results are shown in .

The results demonstrate that the proposed MSP-net achieves high accuracy for eye and mouth state detection. Mouth accuracy is generally higher than eye accuracy because eye state discrimination is affected by interference such as wearing glasses or different glasses types, resulting in slightly lower classification performance. Actual detection effects are shown in [Figure 10: see original paper], with text descriptions in the upper left. The results show MSP-net' s high robustness, accurately recognizing eye and mouth states even when wearing glasses, under low light, and with head rotation.

In addition to high detection accuracy, fatigue detection requires fast processing. shows the computational speed of the three networks for detecting a single 48$×$48 grayscale image on CPU. AlexNet is significantly slower than ResNet and MSP-net due to its much larger number of parameters, making it unsuitable for fatigue detection. MSP-net and ResNet both perform excellently in accuracy and real-time performance, with MSP-net slightly outperforming ResNet.

To prove MSP-Net' s generalizability, comparative experiments were conducted on the ZJU eye state dataset against two other methods. Results are shown in . "Acc." represents accuracy, and "AUC" represents the area under the receiver operating characteristic curve (ROC) [23]. Larger values indicate better model performance. Reference [24] used histogram of oriented gradient (HOG) descriptors and random ferns for eye state classification, while reference [22] proposed a new feature descriptor: Multi-scale Histograms of Principal Oriented Gradients (MultiHPOG) to improve robustness to image noise and scale changes. The results show that our method outperforms both in accuracy and speed.

5.2.2 Fatigue State Detection

Wierwille et al. [11] reported that when entering fatigue, PERCLOS values exceed 0.15, though other researchers have used different thresholds such as 0.25 [12]. This paper considers both PERCLOS and FOM, proposing a fatigue detection scheme: (a) When PERCLOS $\geq$ 0.5 or FOM $\geq$ 0.5, fatigue is detected; (b) If condition (a) is not met, when PERCLOS $\geq$ 0.4 and FOM $\geq$ 0.3 simultaneously, fatigue is also detected.

In fatigue detection system experiments, 10 volunteers simulated driving processes, each simulating fatigue 10 times (100 total fatigue events). Volunteers were instructed to perform various interfering actions such as increased blinking, talking, and light laughing. The system monitored volunteers in real-time using both regular and infrared cameras. records the experimental results.

Precision is calculated as:

$$P = \frac{n_f - n_{mis}}{n_f - n_{mis} + n_{err}}$$

where P is precision, n_f is the number of fatigue events, n_{err} is false detection count, and n_{mis} is missed detection count. Precision reflects false alarm probability—higher precision means lower false detection.

Recall is calculated as:

$$R = \frac{n_f - n_{mis}}{n_f}$$

where R is recall. Recall reflects missed detection probability—higher recall means lower missed detection.

The results show high precision and recall for the system. RGB images from regular cameras achieve higher precision and recall than infrared images simulating nighttime conditions.

5.3 Discussion

Several issues were encountered during experiments. During sample collection, mouth images were captured with non-uniform sizes due to program design. For more effective training, this paper used anisotropic scaling [25] to resize all input images to 48$×$48. In real situations, drivers may talk or laugh lightly, with mouths also opening. However, these openings are much smaller than yawning. To reduce detection errors, samples with slightly open mouths were classified as closed, with only widely open samples considered open.

Initially, eye and mouth state training used binary classification, merging regular and infrared camera samples to detect only eye open/closed and mouth open/closed states. However, results were unsatisfactory, likely because regular and infrared images differ significantly, making it difficult for CNNs to simultaneously accommodate both differences during automatic feature fitting. After analysis, the final approach divided categories into four classes, distinguishing between regular and infrared camera samples.

6 Conclusion

This paper designs a cascaded deep learning structure and a real-time fatigue detection system based on multi-scale pooling convolutional neural networks. First, the MTCNN network detects the driver' s face and extracts key eye and mouth regions. These regions are then fed into the multi-scale pooling MSP-Net for state classification. A fixed-length queue stores detection results per frame within a unit time, and fatigue state is jointly determined through PERCLOS and mouth opening frequency (FOM). Experiments demonstrate that the proposed algorithm achieves high detection accuracy with real-time performance and high robustness to complex environments (such as wearing glasses, moderate head rotation, and nighttime driving). Future work will focus on porting and optimizing the proposed method to embedded platforms.

References

[1] Tefft B C. Acute sleep deprivation and risk of motor vehicle crash involvement [R]. [S.l.]: AAA Foundation for Traffic Safety, 2016.

[2] Viola P, Michael J J. Robust real-time face detection [J]. International Journal of Computer Vision, 2004, 57 (2): 137-154.

[3] Du Yong, Ma Peijun, Su Xiaohong, et al. Driver fatigue detection based on eye state analysis [J]. Technology & Health Care Official Journal of the European Society for Engineering & Medicine, 2008, 23 (s2): S453-S463.

[4] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C]// Proc of International Conference on Neural Information Processing Systems. [S.l.]: Curran Associates Inc, 2012: 1097-1105.

[5] Liu Hailong, Li Baoan, Lyu Xueqiang, et al. Image retrieval based on deep convolutional neural network [J]. Application Research of Computers, 2017, 34 (12): 3816-3819.

[6] Zhu Yu, Zhao Jiangkun, Wang Yining, et al. A review of human action recognition based on deep learning [J]. Acta Automatica Sinica, 2016, 42 (6): 848-857.

[7] Yu Tianshu, Wang Ruisheng. Scene parsing using graph matching on street-view data [M]. [S.l.]: Elsevier Science Inc, 2016.

[8] Shao Hong, Chen Shuang, Zhao Jieyi, et al. Face recognition based on subset selection via metric learning on manifold [J]. Frontiers of Information Technology & Electronic Engineering, 2015, 16 (12): 1386-1396.

[9] Taigman Y, Yang Ming, Ranzato M, et al. DeepFace: closing the gap to human-level performance in face verification [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE Computer Society, 2014: 1701-1708.

[10] Zhang Kaipeng, Zhang Zhanpeng, Li Zhifeng, et al. Joint face detection and alignment using multitask cascaded convolutional networks [J]. IEEE Signal Processing Letters, 2016, 23 (10): 1499-1503.

[11] Wierwille W W, Ellsworth L A. Evaluation of driver drowsiness by trained raters [J]. Accident Analysis & Prevention, 1994, 26 (5): 571-81.

[12] Chu Jiangwei, Jin Lisheng, Tong Bingliang, et al. A monitoring method of driver mouth behavior based on machine vision [C]// Proc of IEEE Intelligent Vehicles Symposium. 2004: 351-356.

[13] Zhao Xuepeng, Meng Chunning, Feng Mingkui, et al. Fatigue detection based on cascade convolutional neural network [J]. Journal of Optoelectronics · laser, 2017 (5): 497-502.

[14] Geng Lei, Yuan Fei, Xiao Zhitao, et al. Driver fatigue detection method based on facial behavior analysis [J]. Computer Engineering, 2018, 44 (1): 274-279.

[15] Yang Shuo, Luo Ping, Chen C L, et al. WIDER FACE: a face detection benchmark [C]// Proc of Computer Vision and Pattern Recognition. [S.l.]: IEEE Press, 2016: 5525-5533.

[16] Jaspers C A M. Histogram equalization: USA, US 6741736 B1 [P].

[17] Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2017, 39 (4): 640-651.

[18] Neubeck A, Gool L V. Efficient non-maximum suppression [C]// Proc of IEEE International Conference on Pattern Recognition. 2006: 850-855.

[19] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[20] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2014, 37 (9): 1904-1916.

[21] Kingma D P, Ba J. Adam: A method for stochastic optimization [C]// Proc of International Conference on Learning Representations. 2015: 1-13.

[22] Song Fengyi, Tan Xiaoyang, Liu Xue, et al. Eyes closeness detection from still images with multi-scale histograms of principal oriented gradients [J]. Pattern Recognition, 2014, 47 (9): 2825-2838.

[23] Fawcett T. An introduction to ROC analysis [J]. Pattern Recognition Letter, 2006, 27 (8): 861-874.

[24] Dong Yanchao, Zhang Yan, Yue Jiguang, et al. Comparison of random forest, random ferns and support vector machine for eye state classification [J]. Multimedia Tools & Applications, 2016, 75 (19): 11763-11783.

[25] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]// Proc of Computer Vision and Pattern Recognition. [S.l.]: IEEE Press, 2014: 580-587.

Note: Figure translations are in progress. See original paper for figures.