Driver Distraction Behavior Detection Method Based on Deep Learning

Peng Mao, Kunlun Zhang*, Da Liang

Key Laboratory of Magnetic Suspension Technology and Maglev Vehicle, Ministry of Education. School of Electrical Engineering

*Corresponding author e-mail: zhangkunlun@home.swjtu.edu.cn

Abstract. With the rapid development of road traffic in China, driver safety accidents caused by road traffic accidents are increasing year by year. According to statistics of relevant departments, 20%-30% of traffic safety accidents are caused by distracted behaviors of drivers. For this reason, this paper proposes a driver distraction behavior detection method based on deep learning, which uses PCN and DSST algorithms for face detection, location and dynamic face tracking. Finally, YOLOV3 object detection algorithm is used to identify distracting behaviors such as smoking and making phone calls around a person's face. The method can detect distracted behaviors in the driving process in real time and has high detection accuracy.

1. Introduction

According to the definition of the International Organization for Standardization (ISO), Driving distraction refers to a phenomenon in which attention is directed to activities that are not related to normal driving, resulting in a decline in driving operation capability. Distracted driving behaviors such as smoking, making phone calls and so on are important causes of traffic accidents [1]. Research on distraction behavior detection technology has become a research focus in the field of rail transit safety [2]. At present, the distraction behavior detection can be divided into detection methods based on traditional techniques and detection methods based on deep learning.

Basing on the traditional technology, the detection method cannot detect the distracting driving behavior such as smoking and calling at the same time, and the two are separately detected. There are mainly monitoring methods based on cell phone signal capture and monitoring methods based on camera to detect the driver's calling behavior during driving. The monitoring method based on cell phone signal capture is to monitor the cell phone signal on the moving vehicle by installing an antenna in a fixed area [3~5], Liu zhi accepts whether the mobile phone from the signal source on the moving vehicle is in a call state by using a directional antenna mounted at a fixed position on the road outside the vehicle [6], Although this method can automatically capture the vehicle that is talking, it is impossible to judge whether the driver or the passenger is talking in the car, so the false detection rate is high. To solve this problem, Ascariz et al. installed the antenna with the signal on the seat in the car, Each antenna has its own analysis circuit. A microcontroller analyzes the signals captured by the two antennas by performing a voltage analysis algorithm to determine whether there is a call behavior and whether the driver or passenger is calling [7], Jie Yang et al. is also a signal-providing method to achieve driver call behavior detection [8]. Based on the camera monitoring method, the driver's call
behavior is analyzed mainly, and the traditional machine learning algorithm is used for evaluation and analysis. The above two traditional detection methods are complicated in structure, slow in detection, and cannot be used in practice. Smoking during driving is through smoke detection, although the method can automatically detect the smoking behavior, but it is impossible to judge whether the driver or passenger smokes, and smoke alarm and other devices are also required. The structure is complex and the false detection rate is high.

The detection method is to realize detection by using convolution network to automatically learn the characteristics of smoking and making phone calls, This paper proposes a new driver distraction behavior detection method based on deep learning. Mainly using PCN (Progressive Calibration Networks) and DSST (Discriminative Scale Space Tracking) algorithms for face detection and dynamic tracking of faces. Then use the trained YOLOV3 (You Only Look Once-V3) model to make a smoking & phone detection in the face area after cutting. The distraction detection model finally obtained has fast detection speed and high accuracy.

2. Proposed Architecture

Deep learning refers to an algorithm set that uses various machine learning algorithms to solve various problems such as images and texts on a multilayer neural network. The core of deep learning is feature learning, which aims to obtain hierarchical feature information through hierarchical networks, thus solving the important problem that required manual feature design in the past. The process of detecting the distraction behavior based on deep learning designed in this paper is shown in Figure 1.

![Distraction detection algorithm flow](image)

**Figure 1. Distraction detection algorithm flow**

2.1. Driver Face Detection

In 2006, Hinton proposed the concept of deep learning and the solution to the problem of gradient disappearance, in 2012 in order to prove the potential of deep learning, the Hinton team participated in the ImageNet image recognition competition for the first time. Through the construction of the CNN (Convolutional Neural Network), Alex Net won the championship and crushed the second place (SVM method) performance [9]. It is also because of this competition that CNN has attracted the attention of many researchers. At present, deep learning has become the hottest research direction in the field of machine learning.

As one of the most important external features of human beings, human faces have also attracted the research interest of many researchers. Every year, excellent face detection algorithms are introduced. This paper mainly adopts the PCN face detection algorithm of Mr. Yamashita. This algorithm is a face detection method based on deep learning. It is mainly realized by using three multi-network cascade methods of face deflection angles calibrated step by step from coarse to fine. Each small step is a simple task based on shallow CNN, which can finally gradually reduce and calibrate the deflected face angle with high accuracy, good effect and less time consumption. The overall network architecture is shown in Figure 2.
Three progressive calibration networks (PCN) to predict face angle values from coarse to fine. PCN-1 performs the 2-class task of face 0 degree and -180 degree, and corrects the face from \([-180°, 180°]\) to \([-90°, 90°]\). PCN-2 performs 3 classification tasks of face 0°, 90°, -90°, and corrects face from \([-90°, 90°]\) to \([-45°, 45°]\). PCN-3 directly performs face angle regression to obtain a rotation angle. The final face angle is PCN-1 angle + PCN-2 angle + PCN-3 angle [10]. As shown in Figure 3.

2.2. Driver Face Tracking
In the driving environment, the background is complex and varied, the driver's face area may move out of the monitoring area, and the facial scale changes, which greatly interferes with the driver's facial features. Although the face detection algorithm based on deep learning PCN has both accuracy and real-time performance, it has high requirements on the hardware components of the image processing unit, which is not conducive to reducing costs. Therefore, based on the actual application scenarios of the system, this paper uses the target tracking method based on Discriminant Scale Space Tracking (DSST) to track the driver's face region in real time and reduce the time consumption of face detection [12].

The DSST algorithm is based on the MOSSE (Minimum Output Sum of Squared Error filter) and adds an improved algorithm after scaling. The core of the MOSSE filter algorithm is to use a filter template and a sequence of images to perform convolution operations to determine the position of the target. The algorithm starts with a set of training images \(f_i\) expected training output \(g_i\), \(g_i\) is usually a Gaussian matrix centered on the target and \(\sigma\) is the variance. The calculation of the filter template \(H\) is as follows:

\[
H_i^* = \frac{G_i}{F_i}
\]

The above formula calculates a filtering template for a single frame of image, but in actual cases, the filtering template needs to be able to adapt to the entire image sequence, MOSSE filtering requires minimal squared error for all actual and desired outputs of the video sequence, which is:

\[
\min_{H^*} \sum_i \left| F_i \circ H^* - G_i \right|^2
\]

Figure 2. The detailed CNN structures of three stages in our proposed PCN method
Interconversion through Parseval's theorem.

\[
H^* = \frac{\sum G_i \ast F_i^*}{\sum F_i \ast F_i^*}
\]

The numerator is the convolution of the input image with the desired input image, while the denominator is the energy spectrum of the input image. After the correlation filter is obtained, inputting the image frame \(Z\) to be detected into the filter, and the correlation score \(y\) is obtained.

\[
y = F^{-1}(H, Z)
\]

**Figure 3.** An overview of our proposed progressive calibration networks (PCN) for rotation-invariant face detection

**Rotate Network** = PCN-1 + PCN-2 + PCN-3

**Figure 4.** DSST algorithm flow

Where \(F^{-1}\) represents the Fourier transform. When \(y\) takes the maximum value, its corresponding position is the update position of the previous frame target in the image frame \(Z\) to be detected. The DSST algorithm designs two consistent correlation filters to achieve target tracking and scale transformation, respectively defined as a translation filter and a scale filter. The former locates the target of the current frame, while the latter estimates the scale of the target of the current frame. The two filters are relatively independent, so different feature types and feature calculation methods can be selected for training and testing. The algorithm schematic is shown in Figure 4.

In this paper, PCN face detection algorithm and DSST video single target tracking algorithm are combined to carry out driver face detection and tracking research. According to the situation of face
occlusion and face rotation, PCN face detection algorithm is used to locate the driver's face region, DSST algorithm is used to dynamically track the face, thus realizing continuous and stable detection and tracking of the driver's face region.

![Figure 5. Face detection effect diagram under smoking](image)

![Figure 6. Face detection effect diagram under the call state](image)

2.3. Distraction Behavior Detection

There are many kinds of distracting behaviors during driving. This article mainly detects the two most common types, smoking and calling. Process A and B have completed the driver's face detection and tracking. Based on this, we cut the face area. Using the current fastest object detection algorithm YOLOV3 to detect the cut region, this scheme detects less background information and clear target location, so the detection speed is faster and the false detection rate is lower.

YOLOV3 series of algorithms solve object detection as a regression problem. Based on a single end-to-end network, the input of the original image to the output of the object position and category is completed. YOLOV3, as the latest algorithm in YOLO series, is Redmon's improved target detection algorithm based on YOLOV2. Different from the Darknet19 network structure used by YOLOV2, YOLOV3 uses the darknet53 network structure [13~14]. As shown in Figure 5. In order to prevent gradient explosion during training, we borrowed from residual network and used a large number of residual blocks in the network.

![Figure 7. YOLOV3_Structure](image)

When YOLOV3 classifies the target object, it does not use Softmax to classify each frame. Because objects appearing in one target box may belong to multiple categories, and Softmax classification can only perform single-class statistics, so YOLOV3 uses more. An independent logistic regression classifier is used for classification. Each classifier only judges whether the object appearing in the target box belongs to the current label, that is, the simple two classification, thus achieving a simple two classification.

YOLOV3 uses the idea of FPN for reference, and adds a series of upsampling modules and convolution modules behind the basic network to generate multi-scale feature maps for detection, so that the model can obtain more image semantic information at lower and higher levels. Each feature
map only predicts 3 bounding box, thus predicting a total of 9 bounding box. The initial prior widths and heights of these 9 bounding boxes are also obtained by clustering.

Using the trained YOLOV3 detection algorithm to detect the distracting behavior such as smoking and calling.

![Smoking & Calling detection](image)

### Figure 8. Smoking & Calling detection

3. Related work
Because PCN face detection algorithm model has been very excellent, especially the detection accuracy of rotating and inclined faces has reached the international leading level, so it has not been retrained. Therefore, the preliminary preparation work is mainly to train and detect the model of drivers smoking and making phone calls. As there is currently no public data set of drivers smoking and making phone calls in the car environment, this paper takes the self-built database as the experimental object, and has collected a total of 30,000 smoking and making phone calls pictures, with a ratio of 1:1. Figure 6 shows a partial image of a self-built database.

![Smoking & Call database](image)

### Figure 9. Smoking & Call database

Due to the differences between the self-built database and COCO and VOC data sets, the K-means clustering algorithm is used to cluster the target frames. After K-means clustering, the model has better learning and expression skills.

| Box Generation | K | AVG IOU | Final Loss |
|----------------|---|---------|------------|
| Cluster IOU    | 9 | 78      | 0.12       |
| Anchor Boxes   | 9 | 65      | 0.33       |

### Table 1. Clustering effect comparison table

4. Performance evaluation
In order to test the detection effect of the proposed algorithm, we mainly use the self-built verification set and the continuous video stream captured by the camera as the test object and respectively use the
algorithm proposed in this paper and YOLOV3 to carry out detection. The detection process is as shown in Figure below.

Figure 10. (d) Ours Method

Figure 11. (e) YOLOV3 Method

The test results of the two are compared as follows.

| Method | Recall | Precision | Speed |
|--------|--------|-----------|-------|
|        |        | Cigarette | Phone |
| YOLOV3 | 80.3%  | 87.2%     | 85.7% | 25FPS |
| Ours   | 85.4%  | 95.7%     | 97.3% | 36FPS |

Comparing the above two methods, we can find that our method has much higher detection accuracy, recall and speed than that of directly using YOLOV3. The main reason is that the method used in this paper only needs to detect the area around the face, with small detection area and less background information. At the same time, after face detection, tracking is added to avoid multiple face detection, so the detection speed and accuracy are high.

5. Conclusion

In this paper, I proposes a method of driver distraction behavior detection based on deep learning. First, this method uses PCN to realize face detection, and then uses DSST to track the detected face. Finally, using YOLOV3 to detect the distraction behavior such as smoking and making phone calls around the face. The experimental results show that the method proposed in this paper has high detection rate and accuracy for distracting behaviors such as smoking, making phone calls, and has strong robustness. Although the method has a high recognition rate, it is easy to detect smoke when objects similar in shape to smoke appear near the mouth. Therefore, the posture estimation of the smoking behavior will be studied to solve the problem of smoke detection by mistake.
Acknowledgments
This work is supported by National Natural Science Foundation (NNSF) of China under Grant 51577160 and the funding of Key Laboratory of Magnetic Suspension Technology and Maglev Vehicle, Ministry of Education .Southwest Jiaotong University

References
[1] Klauer S G,Dingus T A, Neale T A, et al. The impact of driver in-attention on near-erash/crash risk: An analysis using the 100-Car naturalistic driving study data. Washington, D. C: NHTSA, 2006
[2] Kaplan, S. Guvensan M A, Yavuz A G, et al. Driver behavior analysis for safe driving: a survey [J]. IEEE Transactions on Intelligent Transportation on Systems, 2015, 16(6): 1~16
[3] Baojun Liu, Alvarez-Ossa, D., Kherrani, N.P., Zukotynski, S., Chen, K.P, et al. Gamma-Free Smoke and Particle Detector Using Tritiated Foils. Sensors Journal, IEEE, 2007.
[4] M. Abdel-Aty and A. Radwan, Modeling traffic accident occurrence and involvement. Accident Analysis & Prevention, Volume 32, Issue 5
[5] I. Brown, Driver Fatigue. Human Factors: The Journal of the Human Factors and Ergonomics Society June 1994, vol.36, no.2, Pages 298-314
[6] Liu Z, et al. Method and device for automatically obtaining evidence of mobile phone call behavior when driver is driving
[7] J. Ascariz, L. Boquete, J. Cantos and S. Ortega. Automatic system for detecting driver use or mobile phones. Transportation Research Part C: Emerging Technologies, Volume 19, Issue 4, Pages 673-681, 2011.
[8] J. Yang, Simon Sidhom and so on. Detecting Driver Phone Use Leveraging Car Speakers. Proceeding in MobiCom 11 Proceedings of the 17th annual international conference on Mobile computing and networking. Page 97-108, 2011.
[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pp. 1097-1105, 2012
[10] Xuepeng Shi and Shiguang Shan and Meina Kan and Shuzhe Wu and Xilin Chen. Real-Time Rotation-Invariant Face Detection with Progressive Calibration Networks. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Apr 2018
[11] R. Newman Y. Matsumoto et al. Real-Time Stereo Tracking for Head Pose and Gaze Estimation. In Proc of the 4th IEEE International Conference on Automatic Face and Gesture Recognition, Pages 122-128, 2000.
[12] Bolme D S, Beveridge J R, Draper B A, et al. Visual object tracking using adaptive correlation filters [C]. The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR2010, San Francisco, CA, USA, 13-18 June 2010. IEEE, 2010
[13] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. arXiv, 2018
[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. arXiv preprint arXiv: 1506.02640, 2015. 4, 5