Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 13, No. 2, 2024

Research on fatigue driving detection technology based on Dlib and micro-expression

Xinru Fang *, Xinyue Yan

Nanjing Agriculture University, Nanjing 210000, China

* **Corresponding author**: Xinru Fang

Abstract: Driver fatigue is the most important factor contributing to the annual increase in traffic accidents and fatalities. Fatigue can affect driving performance due to lack of concentration and slower reaction time. Therefore, fatigue detection systems are crucial for safe driving. This article proposes a fatigue driving detection scheme based on the dlib library. Build a facial key point detection system based on the Dlib library. Real time calculation of eye and mouth data and head Euler angles using HOG-SVM algorithm, PnP algorithm, and fixed threshold. When running the YOLOV5 model in a PyTorch environment to recognize distracting behaviors, training is conducted on the sample and test sets for recognition, and the accuracy of the model is verified using the validation set for accurate recognition. Train smoking, answering and making phone calls, and drinking water behaviors to achieve distraction detection. Using convolutional neural networks to detect driver status and perform micro expression analysis. Fatigue assessment in multi feature situations of eye, mouth, and head posture and micro expression attention discrimination.

Keywords: Fatigue driving; Facial feature recognition; Micro-expressions; Multi-featured.

1. Introduction

Driver fatigue is one of the main causes of road traffic accidents [1], contributing to approximately 20% of all road traffic accidents. Fatigued drivers exhibit unnatural visible movements such as facial expressions, changing eye movements, percentage of eyes closed (PERCLOS), pupil movements, etc [2].Several studies have used different non-invasive behavioural-based approaches to detect driver fatigue using visual cameras and smart technologies to extract fatigue-related information. Physiological based systems are more reliable and efficient than other techniques because they use physically invasive sensors (electrodes) and sensors to collect information directly from biological signals. Collecting data using physiological methods is a complex and challenging task for researchers, but in recent years the process has become easier with the latest technology. As technology and techniques have advanced, DFD systems have significantly reduced fatigue-related road accidents [3]. To detect the onset of fatigue, the DFD system analyses behavioural and warning signs. The device has the potential to detect early signs of fatigue and reduce the risk of accidents. The system allows operators to visually monitor their level of alertness in real time. Operators can use a variety of techniques to stay alert and manage their fatigue levels in advance. Academics and researchers are using many new technologies and methods to improve the quality of safe driving and to make important decisions in emergency situations [4].

2. Realisation methods

2.1. Facial Feature Recognition

Dlib is a powerful open-source toolkit that provides rich machine learning, deep learning, and image processing capabilities. In the field of computer vision, Dlib is particularly good at face detection and keypoint extraction. Among them, Dlib comes with a 68 keypoint detection model shape_predictor_68_face_landmarks.dat, which can accurately detect the face region from the input image or video and mark the 68 feature keypoints [5].

These 68 keypoints contain rich geometric information about the face, which can be used for subsequent eye and mouth feature extraction and head pose estimation. Specifically, 14 keypoints (shown in Figure 1) can be selected from these 68 keypoints for eye and mouth feature extraction [6]. For example, feature vectors in the region around these keypoints can be extracted using the Histogram of Oriented Gradients (HOG) descriptor, which reflects state information of the eyes and mouth, such as the degree of eye opening, blink frequency, and mouth opening [7].

In addition to eye and mouth features, this keypoint information can also be used to estimate head posture. Using the Perspective-n-Point (PnP) algorithm, the Euler angles of the head can be calculated based on these keypoint coordinates to obtain head rotation and translation information [8]. This head pose information is important in many application scenarios, such as human-computer interaction, virtual make-up testing, motion capture, etc.

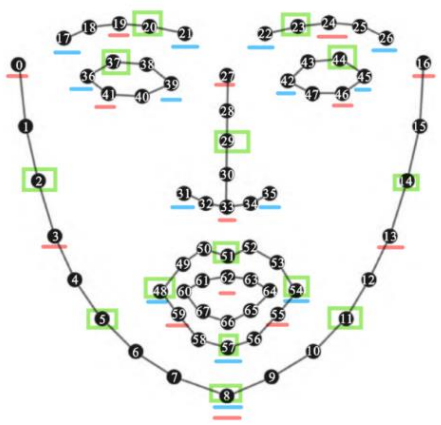


Figure 1. 68 key points in the face 14 feature point selection

Driver fatigue assessment, using eye and mouth data and head Euler angles, can set a fixed threshold or use a machine learning model (e.g. SVM) to judge the driver's fatigue state. The above steps are integrated into a real-time detection system that can continuously detect faces, extract eye and mouth data and head posture, and judge the driver's fatigue state in real-time and provide appropriate feedback.

2.2. Detection of distracting behavior

In this paper, end-to-end convolutional neural network model YOLOv5s model architecture is selected for training and target detection task. Its model architecture consists of four main parts: input, Backbone, Neck, and Head.YOLOv5 is an end-to-end convolutional neural network model for target detection task. It consists of four main parts:Input, Backbone, Neck and Head, which work together to complete the whole process from input image to output detection result. Firstly, the Input receives the original input image and performs some pre-processing operations on it, such as image scale scaling, pixel normalisation, etc. These preprocessing operations help the image data to better fit the model input. Next, the preprocessed image is fed into the Backbone network, which is responsible for extracting features from the image, using a series of improved CSPNet structures and generating multi-scale feature maps. These feature maps contain a variety of features ranging from texture information at the bottom to semantic information at the top. The Neck part then fuses the multiscale feature maps generated by Backbone, using structures such as FPN or PAN to effectively combine features at different scales and generate richer and more representative feature maps.The design of Neck plays a key role in the performance of the final model. Finally, the Head part will perform target detection based on the feature maps output from the Neck.The Head contains multiple detection heads, each corresponding to a feature map of a specific scale. Each detection head predicts the bounding box coordinates, confidence level, and class probability of the target. These outputs are further processed, such as non-extremely large value suppression, confidence threshold screening, etc., to obtain the final detection results.

The main steps of training the YOLOv5 model: The first step is to partition the data set, which is divided into training and validation sets for model training and evaluation. The second step is data augmentation, which performs data augmentation operations on the training set, such as random scaling, flipping, rotating, brightness adjustment, etc., to expand the training data and improve the generalisation of the model. The third step is model initialisation, which initialises the YOLOv5 model using the pre-trained weights. The fourth step is loss function definition, where the GIoU loss function is selected in YOLOv5. The fifth step is model training, where the model is trained using the training set and the loss function is reduced by iteratively optimising the model parameters. The sixth step is model evaluation and tuning, where the validation set is used to evaluate the model, and the model is tuned according to the evaluation results, such as adjusting the hyperparameters and increasing the regularisation.

In this paper, image or video datasets containing mobile phones, water bottles and cigarettes are collected and labelled with images. The location and category of the objects are framed and corresponding labels are defined for distracting behaviours (e.g. using a mobile phone, drinking water and smoking). Use the trained YOLOv5 object recognition and distraction detection model on new images or videos. By passing the input image or video to the YOLOv5 model, the model will output the category of the detected object, the location information and the label for the distracting behaviour.

2.3. Micro-expression recognition

In this paper, a convolutional neural network (CNN) is used to train a classification model for sentiment classification of the extracted features. The deep convolutional neural network consists of three parts, which are convolutional layer, the downsampling layer, and fully connected layer. A camera is used to capture the driver's face image in real time, and fatigue driving detects the driver's facial region, which is subjected to a series of convolutional and downsampling layers to obtain multiple low-dimensional feature vectors. These low-dimensional feature vectors are passed through a fully connected layer, i.e. a traditional neural network input, to obtain the classification results of seven expressions (happy, surprised, focused, confused, bored, sad and tired). According to (previous research), the weights of the seven driving expressions are determined, and the weight assignments of the seven expressions with detailed descriptions are shown in Table 1.

Table 1. Weight assignment for the seven expressions

Emotion Category	Weight	Description
Happy	1.5	Driver is happy.
Surprised	2	Driver is surprised by current road conditions.
Focused	0	The driver doesn't have a lot of emotion.
Confused	-2	Driver doesn't understand current road conditions.
Bored	-1.5	Driver is bored.
Sad	-1	Driver is slightly disgruntled.
Tired	-1	Driver is tired.

Based on the weighting of the above seven expressions, the driver's current concentration score is calculated using the formula shown in equation (1).

$$f(x) = p1 * w1 + p2 * w2 + p3 * w3 + p4 * w4 + p5 * w5 + p6 * w6 + p7 * w7 \quad (1)$$

In order to facilitate the subsequent calculation process, the driver's current concentration score is normalized by controlling the range of score values between 0 and 1 using Equation. Where G(x) is the score after being normalized, f(x) is the score before normalization, and minf(x) and maxf(x) represent the upper and lower limits of the score, with values of -2 and 2 respectively. the normalization formula is shown in Equation (2).

$$G(x) = \frac{f(x) - minf(x)}{maxf(x) - minf(x)} \quad (2)$$

Based on the normalized concentration scores drivers can be classified into three concentration states, namely low concentration state, medium concentration state and high concentration state. This index can reflect the driver's driving state to a certain extent. The score intervals of different concentration degrees are shown in Table 2.

In this paper, the trained expression recognition model is applied to real-time facial video streams to monitor and

identify people's micro-expressions in real time, which is combined with distracted behaviour detection to further determine the driver's fatigue level and provide corresponding emotional feedback and fatigue driving warnings.

Table 2. Different levels of concentration

Level of concentration	Low level concentration	Medium level concentration	High level concentration
Score	0-0.45	0.45-0.75	0.75-1.0

3. The process of fatigue driving detection

Real-time video or image of the driver is captured in real time by camera and other devices, and multi-feature detection is performed using different models. Face key point detection based on the Dlib library to detect the magnitude of changes in the driver's eyes, mouth and nose, and to record the number of blinks and yawns of the driver by continuously monitoring the driver's facial state. Distracted behavior recognition based on the YOLOv5 model, which detects the presence of mobile phones, water cups and cigarettes in real-time video or images, is used to determine whether a driver is engaging in distracting behaviors such as using a mobile phone, drinking water and smoking cigarettes. The number of times a driver uses a mobile phone, drinks water and smokes is recorded by continuously detecting mobile phones, water cups and cigarettes. Based on the CNN expression recognition model for concentration judgement, the probability value of seven types of expressions on the driver's face is detected, and the driver's concentration score is derived by calculation based on the detection results of the seven types of expressions. Multiple features are combined to identify the driver's level of fatigue. If the fatigue level is high, it will force the driver to stop driving, and if the fatigue level is low, it will provide safety tips to the driver.

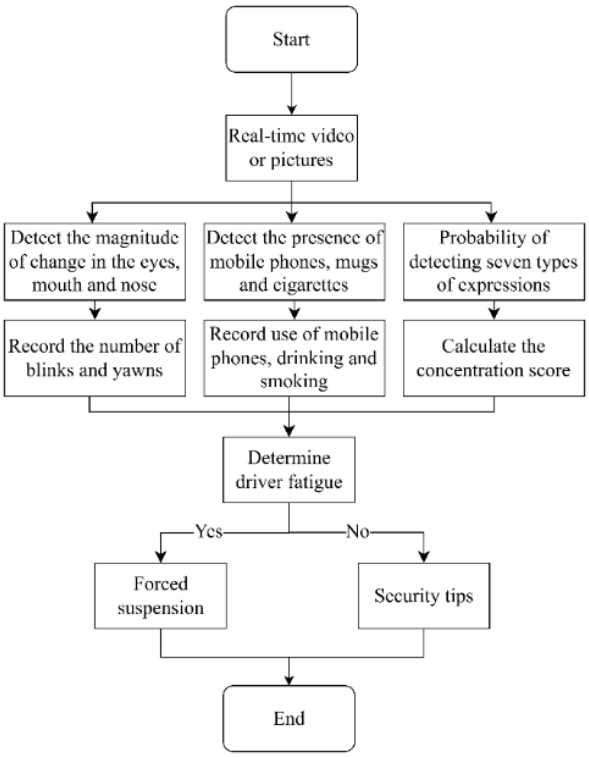


Figure 2. Flowchart of drowsy driving detection

4. Results and analyses

4.1. Facial feature point recognition

The HOG-SVM algorithm is a target detection algorithm that uses Histogram of Orientation Gradients (HOG) descriptors to represent targets in an image, which are then classified by a Support Vector Machine (SVM). In fatigue monitoring, this paper uses the HOG algorithm to extract features from the eyes and mouth, and then feeds these features into the SVM classifier to determine whether the driver is in a fatigued state. The model is evaluated on the validation set after the completion of each epoch in the training process, and the data is stored in a log file with a real-time update window to evaluate the training effect. The recognition accuracy results for the video dataset are shown in Table 3, and the fatigue detection results for the facial feature points are shown in Figure 3.

Table 3. Video dataset recognition accuracy results

Blink detection	Normal condition	Wearing glasses	Strong light	Low light	Long distance + wearing glasses + some non-frontal	Long distance
Before improvement	91%	72%	89%	85%	50%	74%
After improvement	98%	83%	95%	93%	66%	82%

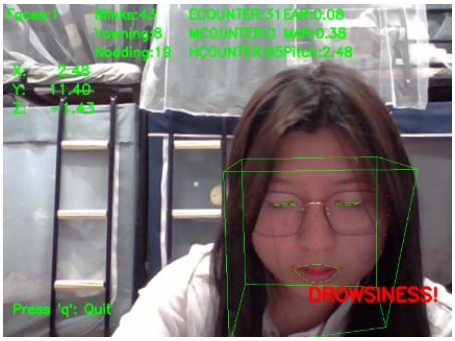


Figure 3. Diagram of fatigue detection results of facial landmark

4.2. Distracted Behavior Recognition

The model has a total of three types of labels, including using mobile phones, drinking water and smoking. The corresponding images are passed to the YOLOv5 model for training, and the training results of the distracting behavior recognition model are shown in Figure 4. First, from the training loss curves, both the training loss and validation loss decrease significantly during the training process and finally converge to a low level, indicating that the model successfully learns the features of distracting behavior and the training process is converged.

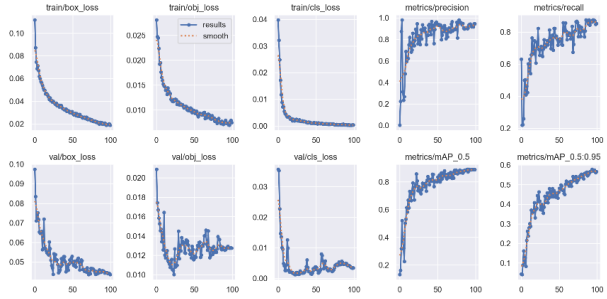
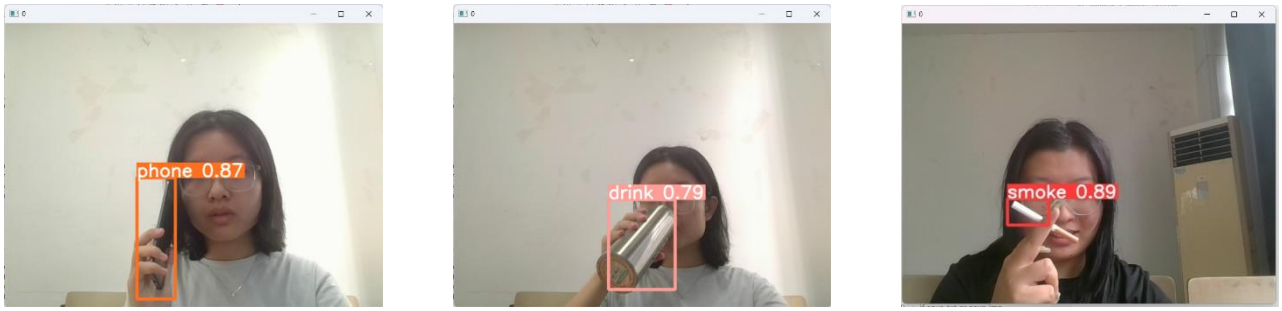


Figure 4. Distraction behavior recognition model training results

As shown in Figure 5, the results of the distracting behavior detection model are presented for scenarios involving mobile phone use, drinking water, and smoking. When the participant is engaged in a phone conversation, the model accurately detects the mobile phone and labels it as "phone" along with the corresponding confidence level, as illustrated in Figure 5a. Similarly, if the participant is drinking water, the model successfully identifies the water cup and labels it as "drink" with the associated confidence level, as depicted in Figure 5b. Furthermore, when the participant is holding a cigarette, the model recognizes the cigarette and marks it as "smoke" with the respective confidence level, as shown in Figure 5c. Overall, the distracting behavior detection model demonstrates robust recognition capabilities across these three typical scenarios, providing valuable input for subsequent attention assessment and behavioral analysis.



a. Calling identification results b. Drinking identification results c. Smoking identification results

Figure 5. Detection results of a distracted behavior recognition model

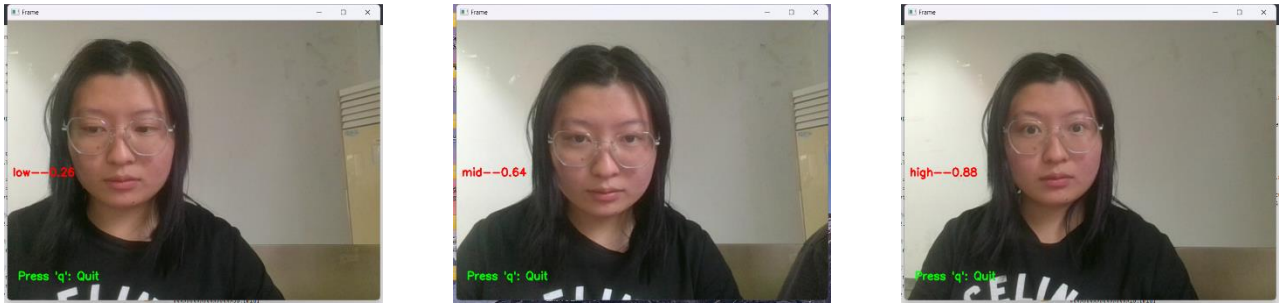
4.3. Micro-expression recognition

When training the classification model using Convolutional Neural Network (CNN) for emotion classification of extracted features, the emotion dataset with labeled emotions is used and the performance of the model is evaluated using techniques such as cross-validation. The training results of the micro-expression recognition model are shown in Figure 6. The loss curve and accuracy curve of the model gradually converge after 50 epochs of training. This indicates that the model has successfully learned the discriminative features of micro-expressions and is able to generalize better to unseen data.

Figure 6. Micro-expression recognition model training results

As shown in Figure 7, the results of the convolutional neural network-based micro-expression recognition model are presented, representing low, medium, and high levels of concentration. When the participant is in a low attention state, the micro-expression recognition model detects this and displays "low" along with the normalized attention score G(x) on the video, as illustrated in Figure 7a. If the participant is in

a medium attention state, the model labels the video with "mid" and the corresponding normalized attention score G(x), as depicted in Figure 7b. Similarly, when the participant is in a high attention state, the model marks the video with "high" and the normalized attention score G(x), as shown in Figure 7c.The convolutional neural network-based micro-expression recognition model has successfully differentiated between the low, medium, and high attention levels of the participants.



a. Low level of concentration b. Medium-level concentration c. High level of concentration

Figure 7. Detection results of a micro-expression recognition model

5. Discuss

Driver fatigue is one of the main risk factors for road accidents every year. Considering physiological and behavioural approaches in combination, physiology-based implementations provide fairly accurate results for fatigue detection but do not facilitate real-time detection, while behavioural features are very effective but not as perfect as biological approaches. Therefore, this study used a combination of biological and behavioural features. This study illustrates how fatigue-related data can be assessed using two different methods: facial feature recognition and micro-expression biosignal processing to assess driver safety. In this experiment, the system uses a video sensor to capture an image of the driver and a bio-signal sensor to analyse the driver's micro-expression signals. To assess driver fatigue. According to the results of the study, the more information used in fatigue detection, the higher the level of performance that can be achieved.

Acknowledgments

This work has been partially supported by the National Innovative Training Programme for University Students (202310307105Z).

References

[1] Xinyun Hu, Gabriel Lodewijks. Detecting fatigue in car drivers and aircraft pilots by using non-invasive measures: The value of differentiation of sleepiness and mental fatigue[J]. Journal of Safety Research,2020,72173-187.

[2] Albadawi Yaman, Takruri Maen, Awad Mohammed. A Review of Recent Developments in Driver Drowsiness Detection Systems[J]. Sensors,2022,22(5):2069-2069.

[3] Abbas Qaisar, Alsheddy Abdullah. Driver Fatigue Detection Systems Using Multi-Sensors, Smartphone, and Cloud-Based Computing Platforms: A Comparative Analysis.[J].Sensors (Basel, Switzerland),2020,21(1):189-201.

[4] Civik E, Yuzgec U. Real-time driver fatigue detection system with deep learning on a low-cost embedded system[J].Microprocessors and microsystems, 2023.

[5] Lashkov I , Kashevnik A , Shilov N ,et al.Driver Dangerous State Detection Based on OpenCV & Dlib Libraries Using Mobile Video Processing[C]//2019 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC).IEEE, 2019.

[6] Rani P I , Muneeswaran K .Recognize the facial emotion in video sequences using eye and mouth temporal Gabor features[J].Multimedia Tools and Applications, 2016, 76(7):1-24.

[7] Angeliki,Fydanaki,Zeno,et al. Evaluating OpenFace: an open-source automatic facial comparison algorithm for forensics.[J].Forensic Sciences Research, 2018.

[8] François Rocca, Mancas M , Gosselin B .Head Pose Estimation by Perspective-n-Point Solution Based on 2D Markerless Face Tracking[C]//International Conference on Intelligent Technologies for Interactive Entertainment. Springer, Cham, 2014.